# Implementation and Evaluation of the Z-Score System for Normalizing Residency Evaluations

Jonathan P. Wanderer, M.D., M.Phil., Getulio R. de Oliveira Filho, M.D., Ph.D., Brian S. Rothman, M.D., Warren S. Sandberg, M.D., Ph.D., Matthew D. McEvoy, M.D.

## ABSTRACT

**Background:** Assessment of clinical competence is essential for residency programs and should be guided by valid, reliable measurements. We implemented Baker's Z-score system, which produces measures of traditional core competency assessments and clinical performance summative scores. Our goal was to validate use of summative scores and estimate the number of evaluations needed for reliable measures.

**Methods:** We performed generalizability studies to estimate the variance components of raw and Z-transformed absolute and peer-relative scores and decision studies to estimate the evaluations needed to produce at least 90% reliable measures for classification and for high-stakes decisions. A subset of evaluations was selected representing residents who were evaluated frequently by faculty who provided the majority of evaluations. Variance components were estimated using ANOVA.

**Results:** Principal component extraction from 8,754 complete evaluations demonstrated that a single factor explained 91 and 85% of variance for absolute and peer-relative scores, respectively. In total, 1,200 evaluations were selected for generalizability and decision studies. The major variance component for all scores was resident interaction with measurement occasions. Variance due to the resident component was strongest with raw scores, where 30 evaluation occasions produced 90% reliable measurements with absolute scores and 58 for peer-relative scores. For Z-transformed scores, 57 evaluation occasions produced 90% reliable measurements with absolute scores and 55 for peer-relative scores. The results were similar for high-stakes decisions.

**Conclusions:** The Baker system produced moderately reliable measures at our institution, suggesting that it may be generalizable to other training programs. Raw absolute scores required few assessment occasions to achieve 90% reliable measurements.
**(Anesthesiology 2018; 128:144-58)**

THE Accreditation Council for Graduate Medical Education (ACGME) has provided a framework of six core competencies for evaluating residents: medical knowledge, patient care, practice-based learning and improvement, professionalism, interpersonal and communication skills, and systems-based practice.[1] These competencies are intended to constitute a system for evaluating residents based on outcomes and performance, but there is no defined evaluation methodology for accurately and reliably assessing these core skills after each rotation. In anesthesiology training, faculty anesthesiologists evaluate resident performance using clinical and professional observations from the immediate perioperative period and other care settings, such as the intensive care unit, preoperative clinic, and pain clinic. However, faculty may have different opinions about acceptable performance,[2] and trainee performance in one situation may not generalize to another.[3] Accordingly, making reliable assessments of resident performance is a challenge and requires multiple points of assessment.

### What We Already Know about This Topic

- Resident evaluations are often idiosyncratic, making it difficult to fairly evaluate both absolute and relative performance
- A previously published system overcomes some of these limitations by converting evaluation metrics into Z scores (deviation from average in SD units), adjusted for faculty grade range use, grade inflation, and resident training level
- The investigators evaluated the system in their residency

### What This Article Tells Us That Is New

- The system was moderately reliable, requiring between 30 and 58 assessments for accuracy
- Fewer assessments were needed with absolute scoring than with peer-relative scoring

Accordingly, residency programs must provide periodic formative, as well as summative, evaluation on all six ACGME core competencies.[1] Developing and interpreting these formative evaluations can be challenging, owing

Supplemental Digital Content is available for this article. Direct URL citations appear in the printed text and are available in both the HTML and PDF versions of this article. Links to the digital files are provided in the HTML text of this article on the Journal's Web site (www.anesthesiology.org).

Submitted for publication June 20, 2017. Accepted for publication September 11, 2017. From the Departments of Anesthesiology (J.P.W., B.S.R., W.S.S., M.D.M.) and Biomedical Informatics (J.P.W.), Vanderbilt University Medical Center, Nashville, Tennessee; and Department of Surgery, Federal University of Santa Catarina, Florianópolis, Brazil (G.R.d.O.F.).






<zdoi;10.1097/ALN.0000000000001919>

to evaluation biases that include grade inflation and idiosyncratic grade-range usage. Baker[4] described a system that normalized resident evaluations to adjust for variations in individual faculty anesthesiologist assessments, idiosyncratic grade-range usage, and resident level of training. This system consists of an instrument to measure peer-relative (relative-to-peer) and absolute (anchored) performance in all six core competencies, as well as Z-score transformations that convert the raw measurement data from the assessment tool into normalized Z scores. This system is intended to be used in transforming "noisy" evaluation data into valid and reliable signals that can be utilized to rank-order residents by performance and identify residents that need an intervention to address performance difficulties in the core competencies.

The Z-transformed scores combine and normalize the instrument's Likert items into standard scores. These standard scores set the mean score to 0 and represent all scores in terms of SDs above or below the mean score. Thus, a Z-transformed score of –0.5 would represent a score that is half a SD below the mean. Baker grouped the peer-relative scores together into one combined measurement ($Z_{rel}$) and the absolute scores together into another combined measurement ($Z_{abs}$). Thus, the system relies primarily on two scores that are cumulative in nature rather than individual instrument items, with these scores being able to be used in formative or summative assessments and feedback. Baker proposed thresholds for residents in need of intervention, those experiencing a challenge, and those facing serious performance issues.[5]

In July 2012, our institution implemented an evaluation system identical to Baker's, with absolute and peer-relative measurements that represent the six core competencies, as well as flags for concerns about essential competency attributes, faculty confidence assessment, and free text comments. Absolute measurements are comparisons to fixed competency standards (*i.e.*, criterion-based), whereas peer-relative measurements are comparisons to peers (*i.e.*, norm-referenced), as described further below. Although the system described by Baker was based on large sample size of 14,469 evaluations, it was also performed at a single center, and it is unclear whether the findings and methodology would be applicable at another institution or whether the utilization of summative scores is justified. Additionally, the number of measurement occasions (clinical encounters with subsequent evaluation) necessary for dependable measurements was not defined. Thus, we undertook an implementation, validation, and analysis of the Baker Z-score system at our institution. Our study was conceived as a planned attempt at reproduction of a notable finding in a domain of educational research for valid and reliable trainee assessment.

## Materials and Methods

This study was deemed exempt by the Human Research Protection Program/Institutional Review Board of Vanderbilt University (protocol 130507), as it was research conducted in an existing educational setting. At our anesthesiology residency training program, faculty anesthesiologists are assigned one evaluation per week for each resident they supervise. Our institution uses a web-based platform provided by New Innovations (New Innovations, Inc., USA) to solicit, store, and aggregate residency evaluations. An evaluation instrument containing absolute and peer-relative measurements identical to that described by Baker was used during the time period evaluated (appendix 1). Using this instrument through the New Innovations platform, faculty anesthesiologists may create and enter resident evaluations at any time with a minimum of one evaluation per resident per week requested from each faculty member with whom the resident worked in a given week (Monday through Sunday). Thus, residents receive more evaluations when working with several faculty members in a week than when working with the same faculty member for a week. As a result, residents are evaluated more frequently while on operating room-based rotations compared to non–operating room–based rotations, such as pain medicine clinic or critical care medicine rotations. At the end of each week after the evaluations were assigned, faculty members received an email reminder if they had not completed all assigned resident evaluations.

These evaluations were transmitted in an aggregated file by secure file transport protocol on a monthly basis. Once an evaluation file was received, it was processed by an automated system task that updated a local SQL Server (Microsoft, USA) database. The resulting Z-scores and deidentified raw evaluation data were provided to one of the authors (J.P.W.), who performed manual validation. Comparison of the results from the initial Z-score SQL Server implementation and manual review revealed two significant interpretation discrepancies. Specifically, the original description of the methodology was ambiguous as to how scores should be aggregated for both the relative and absolute measurements and to which set of data means should be applied.[5] The technical implementation was applied as a mean of the competencies aggregated by means, *i.e.*, a mean of means, whereas manual validation was performed by taking a mean of all of the individual competency scores. The differing interpretations of the original methodology were resolved after discussion with Dr. Keith H. Baker, M.D., Ph.D., Department of Anesthesia, Critical Care and Pain Medicine, Massachusetts General Hospital, Boston, Massachusetts (personal communication, November 2012), resulting in our final SQL Server implementation (appendix 2, Supplemental Digital Content, http://links.lww.com/ALN/B546). For analysis, we extracted raw scores from this system between July 2012 and June 2015.

### Generalizability and Decision Studies

Generalizability studies examine the dependability of behavioral measurements, taking into consideration the magnitude of the multiple sources of measurement error imposed by the situations under which measures were obtained—the universe

of generalization. When assessing learner achievement, such situations—facets—usually are represented by item characteristics, rater biases, measurement occasions, and evaluation designs. The error variance of each of these components affects the resulting measure of persons' behavior that is the object of measurement. By identifying the sources of error and their respective size—the percentage of total error variance—decision makers may identify the factors that contribute to the dependability of performance measures. Decision studies use the results of generalizability studies to inform decisions about changes in the universe of observation.

Four generalizability studies were conducted to assess the reliability of resident measures produced by raw absolute, raw peer-relative, Z-transformed absolute, and Z-transformed peer-relative scores. For these studies, we created a balanced sample from the original data set, using the criterion of 15 evaluations per resident provided by different faculty. This best represented the most frequent situation in our actual environment: one random faculty providing an evaluation of one resident only once, representing 49% of faculty/resident interactions. We chose 15 evaluations per resident because that was the harmonic mean of the number of evaluations per resident in the original data set.[5] Our balanced sampling strategy produced data from 80 resident (person facet) evaluations (78% of the number of residents in the original data set) by 78 faculty members (rater facet) (55% of the faculty evaluators in the original data set) on 15 unique weekly evaluations (occasion facet), resulting in 1,200 unique resident evaluations available for analyses. This sample size exceeds the 50 to 500-observation samples adequate to produce robust generalizability studies.[6,7]

To assess potential raters × item interactions, two generalizability studies were performed on the scores of the items of the evaluation instrument (item). These studies included items, raters, and occasions as potential sources of error variance on persons' scores. The P × R:O × I design (persons crossed with raters within occasions crossed with items) was chosen based on the assumptions that all persons had been rated the same number of times (15 occasions) by one different rater on each occasion on the same seven items that comprised each scale. Items were considered fixed, and all other facets were random. Individual item reliability estimators were extracted from these generalizability studies.

Once assured the absence of rater × item interaction (variance component = 0), four generalizability studies were conducted to investigate the reliability of person measures obtained with raw relative and absolute scores and their counterpart Z-transformed scores. For these studies, we chose a partially nested design P × R:O (persons crossed with raters within occasions), because all persons were evaluated the same number of times by one different rater on each measurement occasion. From these studies, decision studies were performed to estimate the number of occasions one resident should have to be evaluated to produce 90% reliable performance measures. Using data from generalizability studies on Z-scores, the dependability of the thresholds proposed by Baker for suboptimal performance (−0.5, −0.6, and −0.8 SD) was investigated.

To assess the variance components of the rater × occasion interaction and their impact on the reliability coefficients, further generalizability studies were conducted using a P:R × O (persons-nested-within-raters-crossed with occasions) design. For these studies, a random balanced sample containing 1,040 data units was drawn from the original data set. The sample comprised evaluations performed by 52 faculty raters of 10 different residents each, on two occasions for each resident (52 × 10 × 2). Generalizability studies were performed on raw and Z-transformed relative and absolute scores.

Scores in the samples used for generalizability studies were compared with scores of data units not used for the generalizability studies by Student's *t* tests for independent samples to assure that the samples used for generalizability studies were representative of the data set regarding the summative scores. A two-sided *P* value of 0.05 represented statistical significance. EduG software (Swiss Society for Research in Education Working Group, Switzerland) was used to perform the analysis.

## Results

From July 2012 to June 2015, 10,525 evaluations were identified for analysis. After discarding incomplete evaluations, 8,754 evaluations remained. These evaluations were entered by 141 faculty members for 102 residents (CA-1 = 250; CA-2 = 3,456; CA-3 = 2,983; and CA-4 = 2,065). The number of evaluations per faculty member ranged from 1 to 349, with a median of 42 evaluations. The number of evaluations received by residents ranged from 1 to 203, with a median of 83 evaluations.

Factor analysis with principal component extraction and orthogonal rotation identified a single factor in each scale. For the absolute scale, the Eigenvalue was 6.33 with 91% explained variance. For the peer-relative scale, the Eigenvalue was 5.95 with 85% explained variance. Given the presence of a single dominant factor, scores for generalizability studies were estimated as the average of the item scores.

### *Generalizability and Dependability*

As described above, 1,200 evaluations were selected for generalizability and decision studies (fig. 1). Out of the four types of scores analyzed, raw absolute scores had the highest degree of variance due to differences between residents (23.2%), followed by raw peer-relative scores (15.9%), Z-transformed peer-relative scores (14.5%), and finally Z-transformed absolute scores (13.8%), as described in table 1. Z-transformed scores had higher standard errors compared to raw scores, relative to score means (table 2). Variance due exclusively to differences between residents was more strongly captured by raw scores compared to the Z-transformed scores. We noted that the greater variance in the persons facet, the greater the reliability of the measures, as can be observed in the respective generalizability

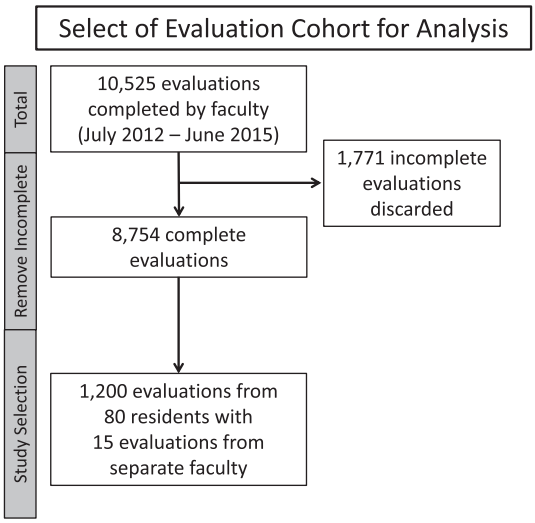



**Fig. 1.** This image depicts the cohort selection. Incomplete evaluations were removed before performing the principle component analysis, and a subset of complete evaluations was selected for generalizability and decision studies.

coefficients. Measurement occasions *per se* did not contribute substantively to the error variance. The major component of score variance was the interaction between persons and measurement occasions.

Based on the decision studies, the estimated number of evaluations needed to produce 90% reliable measures for classification purposes were estimated as 30 for raw absolute scores, 47 for raw peer-relative, 57 for Z-transformed absolute, and 55 for Z-transformed relative scores. Similar figures were estimated for 90% reliable absolute (summative) decisions: 30 for raw absolute scores, 48 for raw peer-relative, 57 for Z-transformed absolute, and 54 for Z-transformed relative scores (figs. 2 and 3). Phi coefficients (dependability indexes) were estimated for high-stakes decisions based on the thresholds Z-scores defined in Baker's studies at −0.8, −0.6, and −0.5 SD. High dependability (reliability) of decisions based on these thresholds was predicted (table 3).

Table 4 shows the results of the generalizability studies designed to disclose the amount of score variance due to rater x occasion interactions. The values varied from 0.2 through 0.6% of total variance, suggesting that faculty were consistent in their ratings across measurement occasions. The variance attributable to differences among raters was apparent only for raw scores. As expected, Z scores were not affected by differences among raters' rating styles or preferences. However, greater residual error variance was found in Z scores, resulting in lower reliability, as indicated by their respective generalizability coefficients and standard errors.

**Table 1.** Contribution of Residents, Measurement Occasions, and Faculty Raters on Score Variance

| | Raw Absolute | | Raw Peer-relative | | Z-transformed Absolute | | Z-transformed Peer-relative | |
|---|---|---|---|---|---|---|---|---|
| Source | Component | % | Component | % | Component | % | Component | % |
| Residents (P) | 0.293 | 23.2 | 0.069 | 15.9 | 0.148 | 13.8 | 0.136 | 14.5 |
| Measurement occasion (O) | 0.022 | 1.8 | 0.012 | 2.8 | 0.003 | 0.3 | 0.002 | 0.2 |
| Faculty rater providing evaluation at each clinical encounter (R:O) | — | — | — | — | — | — | — | — |
| Resident physician with measurement occasions (P:O) | 0.948 | 75.0 | 0.352 | 81.3 | 0.927 | 86.0 | 0.798 | 85.3 |
| P:R:O (error) | — | — | — | — | — | — | — | — |
| Total | | 100 | | 100 | | 100 | | 100 |

Measurement occasion = clinical encounter of faculty with resident; O = occasion; P = persons (residents); R = raters (faculty).

**Table 2.** Generalizability Coefficients and Errors of Measurement

| Score | Differentiation Variance | Relative Variance | Absolute Variance | Grand Mean Score | SE of the Grand Mean | −95% CI of Mean Score | +95% CI of Mean Score | G-relative | G-absolute |
|---|---|---|---|---|---|---|---|---|---|
| Raw absolute | 0.29 | 0.06 | 0.06 | 4.90 | 0.08 | 4.75 | 5.05 | 0.82 | 0.65 |
| Raw peer-relative | 0.07 | 0.02 | 0.02 | 3.64 | 0.04 | 3.55 | 3.72 | 0.75 | 0.75 |
| Z-transformed absolute | 0.14 | 0.06 | 0.06 | 0.01 | 0.05 | −0.10 | 0.11 | 0.70 | 0.70 |
| Z-transformed peer-relative | 0.14 | 0.05 | 0.05 | −0.02 | 0.05 | −0.12 | 0.08 | 0.72 | 0.72 |

The greater the SE, the greater the expected measurement error affecting the global scores. The percentage of SE is calculated as (SE/grand mean) × 100, the size of the measurement error relative to the grand mean expressed as a percentage.

Absolute variance = sum of variance components of all facets included in the universe of admissible observations; +/−95% CI mean score = upper and lower 95% CIs of the grand mean; Differentiation variance = variance attributed to the object of measurement (residents); G-absolute = generalizability coefficient for absolute decisions = differentiation variance/(differentiation variance + absolute variance); Grand mean score = mean score across all facets of the universe of admissible observations = overall mean score; G-relative = generalizability coefficient for relative (classification) decisions = differentiation variance/(differentiation variance + relative variance); Relative variance = sum of variances of components that include the object of measurement facet (interaction or nesting); SE of the grand mean = the square root of the variance of the grand mean (measure of accuracy of scores used to estimate CIs around the grand mean).

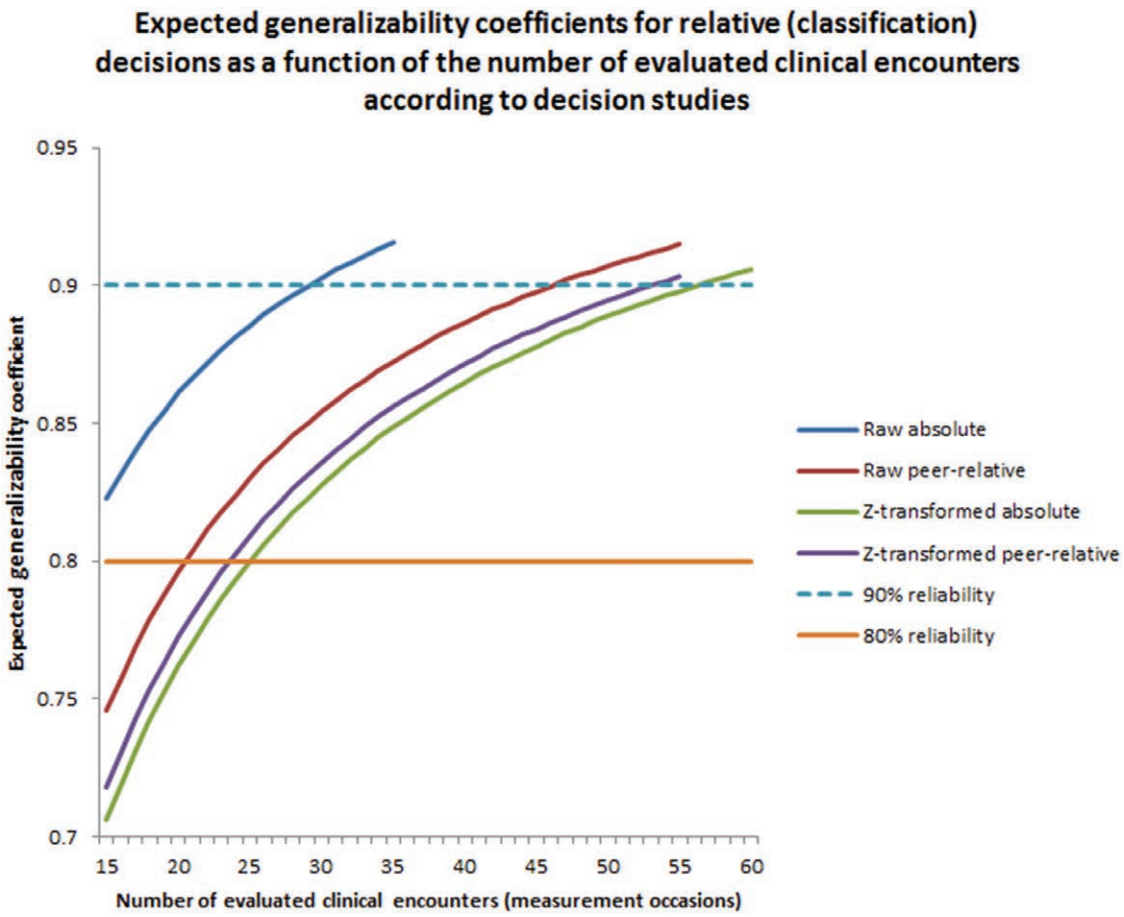



**Fig. 2.** This image summarizes decision studies' predictions of generalizability coefficients for relative (classification) decisions according to the number of clinical encounters. To facilitate visualization, 90 and 80% reliability *horizontal lines* are included. For example, raw absolute scores are expected to produce 90% reliable assessments after 30 evaluated clinical encounters, whereas raw peer-relative scores are expected to produce 90% reliable assessments after 47 evaluated clinical encounters. In comparison, Z-transformed absolute scores are expected to produce 90% reliable assessments after 57 evaluated clinical encounters, whereas Z-transformed peer-relative scores are expected to produce 90% reliable assessments after 55 evaluated clinical encounters.

Raw absolute scores were significantly higher in the sample used for the P × R:O × I and P × R:O studies as compared to the remaining data set. In the sample used for the P:R × O generalizability study, the raw relative scores were significantly higher than those of the remaining data set. No other differences were observed between study samples and the remaining data set (table 5).

## Discussion

Having a valid, reliable, quantitative, and stable measurement of resident training performance is crucial for informing decisions regarding professional development, promotion, and, when needed, remediation.[8] Along with robust conventional mechanisms, including informal feedback by faculty and free text comments, these measurements could potentially identify those in need of remediation and could serve as an early warning system for others.[5] Accordingly, we have described the technical aspects of a real-world implementation of the Baker Z-score system in a large residency program (18 residents/yr). Our study has four important findings that add to the literature on resident evaluation. First, we performed a psychometric analysis of the assessment instruments in the Baker evaluation system and demonstrated that the raw scores account for the variance between persons being rated (residents) better than the Z-transformed scores, which was unexpected. The variance attributed to differences among residents was smaller than that associated with the interaction between residents and measurement occasions, indicating that the scores attributed to the residents were homogenous but varied across measurement occasions. Second, our work demonstrates that the Baker evaluation system using raw or Z-transformed scores produces reliable scores and confirms that it is appropriate to use in formative and summative assessments for measures of resident performance. Third, we demonstrated that fewer rating occasions are needed to reach 90% reliability of the scores produced when using raw absolute scores as compared

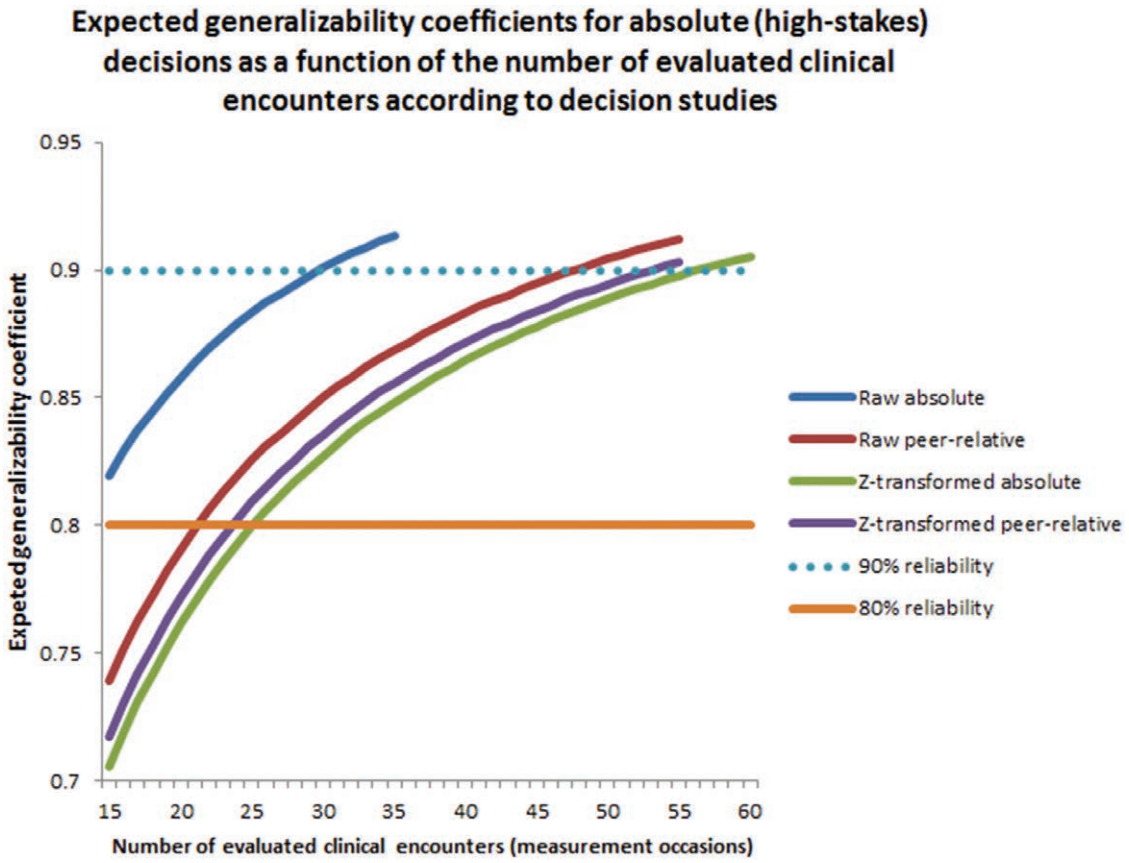



**Fig. 3.** This image summarizes decision studies' predictions of generalizability coefficients for absolute (high-stakes) decisions according to the number of clinical encounters. To facilitate visualization, 90 and 80% reliability *horizontal lines* are included. For example, raw absolute scores are expected to produce 90% reliable assessments after 30 evaluated clinical encounters, whereas raw peer-relative scores are expected to produce 90% reliable assessments after 48 evaluated clinical encounters. In comparison, Z-transformed absolute scores are expected to produce 90% reliable assessments after 57 evaluated clinical encounters, whereas Z-transformed peer-relative scores are expected to produce 90% reliable assessments after 54 evaluated clinical encounters.

**Table 3.** Phi Coefficients at Baker's Threshold Z Scores

| | Z-score = −0.8 | Z-score = −0.6 | Z-score = −0.5 |
|---|---|---|---|
| Z-transformed absolute | 0.92 | 0.89 | 0.86 |
| Z-transformed peer-relative | 0.93 | 0.90 | 0.87 |

to Z scores. Finally, we demonstrated high dependability of the Z-transformed score thresholds identified by Baker, which could be readily operationalized by a clinical competency committee using these data as part of a structured assessment process.

A key finding of Baker's study was "the low correlation between first and second $Z_{rel}$ scores when a faculty member evaluated the same resident on two occasions."[5] The author concluded that "a single $Z_{rel}$ score has only a small amount of clinical performance 'truth' associated with it." This finding matches ours, justifies the approach used in generalizability analysis of choosing a single rater for each measurement occasion, and is consistent with the large amount of variance associated with the interaction of persons and measurement occasions. Baker justified his finding by invoking the context sensitivity theory. Our findings, achieved with different raters in each measurement occasion, also could be explained by context sensitivity. Both studies agree that reliability is dependent on multiple measurement occasions. Our study went further in estimating how many occasions and the amount of consistency of measures depending on the number of measurement occasions. We cannot compare our results regarding raw scores, because they were not analyzed in Baker's original study. However, greater reliability was found for raw scores compared to Z-transformed scores. This was caused by the greater measurement error associated with Z-transformed scores.

Both generalizability studies presented in this manuscript show that the interactions between residents and measurement occasions contribute with substantive amounts for the error variance of scores produced by Baker's evaluation system. We have also shown that the rater per measurement occasions of the same resident is highly consistent, as the negligible amounts of variance associated with such

**Table 4.** Components of Variance and Generalizability Coefficients from P:R×O Generalizability Studies

| Source | Raw Absolute, % | Raw Peer-relative, % | Z-transformed Absolute, % | Z-transformed Peer-relative, % |
|---|---|---|---|---|
| P:R | 24.5 | 29.7 | 33.0 | 45.1 |
| R | 58.2 | 43.6 | 0.0 | 0.0 |
| O | 1.7 | 1.0 | 1.2 | 0.4 |
| RO | 0.3 | 0.2 | 0.4 | 0.6 |
| PO:R | 15.3 | 25.5 | 65.5 | 53.9 |
| Generalizability coefficients absolute and relative decisions | 0.91 | 0.85 | 0.50 | 0.62 |
| SE absolute decisions | 0.30 | 0.23 | 0.59 | 0.52 |
| SE relative decisions | 0.31 | 0.24 | 0.59 | 0.52 |

O = occasion; P = persons (residents); PO:R = residual error variance; P:R = Residents evaluated by each faculty, represent the object of measurement; R = raters (faculty); RO = error variance of scores caused by interaction between faculty and occasions).

**Table 5.** Comparison of Samples Used in the Generalizability Studies *versus* Those Not Selected for Inclusion

| | Raw Absolute | | Raw Peer-relative | | Z-transformed Absolute | | Z-transformed Peer-relative | |
|---|---|---|---|---|---|---|---|---|
| | Out | In | Out | In | Out | In | Out | In |
| Sample 1, mean (SD) | 4.79 (1.02) | 4.9 (1.12) | 3.63 (0.72) | 3.64 (0.66) | 0 (0.97) | 0.01 (1.04) | 0 (0.98) | –0.02 (0.97) |
| Sample 1, mean difference [95% CI] | –0.1 [–0.17 to –0.03]* | | –0.01 [–0.05 to 0.03] | | –0.01 [–0.07 to 0.05] | | 0.02 [–0.04 to 0.08] | |
| Sample 2, mean (SD) | 4.8 (1.03) | 4.85 (1.06) | 3.64 (0.72) | 3.56 (0.65) | 0 (0.97) | 0 (1.02) | 0 (0.98) | 0 (1) |
| Sample 2, mean difference [95% CI] | –0.04 [–0.11 to 0.02] | | 0.07 [0.03 to 0.12]† | | 0 [–0.07 to 0.06] | | –0.01 [–0.07 to 0.06] | |

*$P$ = 0.003. †$P$ = 0.001.

In = included in generalizability studies; Out = not included in generalizability studies; Sample 1 = sample used for the P × R:O × I and P × R:O studies; Sample 2 = sample used for the P:R × O generalizability study.

interaction suggest. This occurs in the presence of high heterogeneity among raters in attributing raw scores, as the substantive amount of error variance associated with the rater facet suggests. Put together, our results are consistent with the conclusion that Baker's system captures differences in situation-specific resident performance. This is highly desirable, because resident performance is expected to remain unstable—to fluctuate—during the learning curves of complex anesthetic procedures.[9]

A final point of clarification for interpreting the results of this study is that by nesting only one rater within each measurement occasion, we were unable to estimate variance for the rater within the occasion facet. This was necessary to explore the effect of measurement occasions. For this reason, we conducted another set of generalizability studies designed to explore the consistency of raters' scoring across repeated occasions. The negligible amount of variance associated with rater × occasion interactions suggests that raters are consistent in their raters of each resident in at least two consecutive measurement occasions. Such behavior applies to raw and Z scores.

To put our results into practical context, the level of reliability desired for making formative or summative (high-stakes) decisions should be understood. High-stakes exams (*e.g.*, licensing board exams) have a goal of at least 90% reliability, whereas formative assessments accept anything more than 70% as being sufficient.[10,11] The decision studies in our analysis demonstrated that when using raw absolute scores, having 30 evaluations per resident would produce 90% reliability, whereas 15 evaluations would produce a reliability of 82% (fig. 3). As a practical example, if an anesthesiology resident received two faculty evaluations per week of work in the operating room, then the compilation of evaluation scores for consideration by the clinical competency committee and for use by the program director in their quarterly review (formative assessment) would include 24 evaluations and have greater than 80% reliability concerning statements made about their performance if the raw scores are used. If this is extended to the 6-month evaluation period with input required for high-stakes reporting to the ACGME and American Board of Anesthesiology, a resident would have 48 evaluations, and both raw absolute or raw peer-relative scores would produce greater than 90% reliability (fig. 3). Of note, raw and Z-transformed absolute and peer-relative scores produce adequate reliability to be used for formative assessment if more than 15 evaluations are completed on the trainee (fig. 2). This level of reliability in formative and summative assessments could be of great assistance to educators in making decisions about resident progression or remediation throughout training.

Of note, this current project was implemented before the start of the ACGME Milestone era for anesthesiology. In the

Milestone system, which now delineates 25 subcompetencies spread among the 6 core competencies, individual absolute rankings are more desired than peer-relative or training year-relative rankings, because the goal is individual progression toward unsupervised practice with recognition that trainees may progress on different learning trajectories.[1] This finding may be of particular importance to training programs across the country, because program directors have been given no specific direction on how to implement evaluation schemas in the Milestone era, and numerous questions remain and are debated. For instance, should the subcompetencies be used verbatim as an assessment tool? Or should an alternative evaluation system be used that the clinical competency committees and program directors use to map to the Milestones system for reporting? Based upon our results, use of absolute scoring within the five levels of the Milestones rubric is still needed, but peer-relative assessments could be abandoned in favor of absolute ranking scales focused on the individual learner. Although these recommendations can be made based upon our results, further psychometric evaluation should be undertaken to ensure that reliability of the scores does maintain when using Milestone rankings.

Finally, external validation of competency assessment tools, such as this study, are important in testing whether original research findings are robust to generalization to other settings. As described above under Materials and Methods, unintentional ambiguity in initial descriptions of systems can lead to erroneous implementation if not carefully checked during implementation. Providing reproducible work in the form of shared code, as we have done (appendix 2, Supplemental Digital Content, http://links.lww.com/ALN/B546), can reduce these risks.

The present study does have several limitations. First, because the evaluation system utilized before 2012 by our residency program had a different set of questions, we were unable to evaluate the Z-score system on our historical evaluation data collected before that time to account for historical trends. Second, we did not analyze all aspects of the system that Baker described, omitting analysis of the case confidence scores, essential competency attributes, and qualitative assessment of free text comments. This approach was in large part tactical, because we were attempting to determine the feasibility and appropriateness of incorporating a summary metric into our clinical competency committee process and have modified our case confidence scores from Baker's description to fit our rotations in a more specific manner. Third, generalizability theory deals basically with random effects. For our generalizability studies, random facets were created by randomly sampling from the original database levels of each facet included in the study according to the intended study design. Therefore, the random nature of our balanced samples does not imply that data were collected from faculty following any random assignment scheme. Finally, our results were from a single institution, which may impact their generalizability. Our approach to discussing education and evaluating residents likely has important differences compared to other institutions, which may have influenced our results.

The future of education research should include studies that determine the predictive value of assessment scores in identifying residents who will have difficulty during residency training. One additional study, for instance, would be investigating the relationship between our evaluation data and educational outcomes, such as clinical competency committee referrals and board examination results, which would complement the analysis described above. These outcomes have been positively linked elsewhere.[12] Additionally, future studies need to determine whether assessment data can be leveraged to facilitate clinical education within an anesthesiology residency program. For instance, we have described a decision support system for resident operating room assignments, which provided summaries of resident case experience to assist with creating appropriate clinical assignments.[13] Highly reliable assessment scores generated by a valid scoring system could be incorporated into such a decision support system, providing information to faculty about ongoing assessments of trainee competence, in addition to simply case numbers performed. However, because absolute scores had higher reliability, the next step would be to perform a standard setting procedure (*e.g.* Angoff) to create criteria for passing and failing.

In conclusion, we report on the implementation and external validation of a resident assessment tool. We determined that the Baker assessment system produces moderately reliable measures from a reasonable number of measurement occasions such that formative and summative assessment decisions can be made, with raw absolute scores requiring the fewest measurement occasions for comparable reliability.

## Acknowledgment

The authors thank Nimesh Patel for his efforts in developing our SQL implementation of the Z-score system.

## Research Support

This work was supported by the Department of Anesthesiology, Vanderbilt University Medical Center, Nashville, Tennessee. Dr. Wanderer was funded by the Foundation for Anesthesia Education and Research and the Anesthesia Quality Institute's Mentored Research Training Grant-Health Services Research.

## Competing Interests

The authors declare no competing interests. Dr. McEvoy received funding (not related to this article) from the GE Foundation for educational research work in Kenya, from Edwards Lifesciences for research in goal-directed fluid therapy, and from Cheetah Medical for research in goal-directed fluid therapy.

## Correspondence

Address correspondence to Dr. Wanderer: Vanderbilt University Medical Center, 1301 Medical Center Drive, TVC 4648, Nashville, Tennessee 37204. jonathan.p.wanderer@vanderbilt.edu. Informa-

tion on purchasing reprints may be found at www.anesthesiology.org or on the masthead page at the beginning of this issue. ANESTHESIOLOGY's articles are made freely accessible to all readers, for personal use only, 6 months from the cover date of the issue.

## References


1. Bell HS: How should the ACGME core competencies be measured? Acad Med 2009; 84:1173; author reply 1173
2. Downing SM: Threats to the validity of clinical teaching assessments: What about rater error? Med Educ 2005; 39:353–5
3. 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(suppl 10):S56–60
4. Baker K: Determining resident clinical performance: Getting beyond the noise. ANESTHESIOLOGY 2011; 115:862–78
5. Brennan, RL: Generalizability Theory. New York, Springer-Verlag, 2001, p. 381
6. Atilgan H: Sample size for estimation of g and phi coefficients in generalizability theory. Eurasian J Educ Res 2013; 51:215–28
7. Maas CJM, Hox JJ: Sufficient sample sizes for multilevel modeling. Methodology 2005; 1:86–92
8. Nabors C, Forman L, Peterson SJ, Gennarelli M, Aronow WS, DeLorenzo L, Chandy D, Ahn C, Sule S, Stallings GW, Khera S, Palaniswamy C, Frishman WH: Milestones: A rapid assessment method for the Clinical Competency Committee. Arch Med Sci 2017; 13:201–9
9. de Oliveira Filho GR: The construction of learning curves for basic skills in anesthetic procedures: An application for the cumulative sum method. Anesth Analg 2002; 95:411–6
10. Downing SM: Reliability: On the reproducibility of assessment data. Med Educ 2004; 38:1006–12
11. Wainer H, Thissen D: How is reliability related to the quality of test scores?: What is the effect of local dependence on reliability? Educ. Measurement Issues Practice 1996; 15:22–9
12. Baker K, Sun H, Harman A, Poon KT, Rathmell JP: Clinical performance scores are independently associated with the American Board of Anesthesiology certification examination scores. Anesth Analg 2016; 122:1992–9
13. Wanderer JP, Charnin J, Driscoll WD, Bailin MT, Baker K: Decision support using anesthesia information management system records and accreditation council for graduate medical education case logs for resident operating room assignments. Anesth Analg 2013; 117:494–9

## Appendix 1

**Anesthesiology Resident Evaluation**

PRG 2
**DA:ANES:MSA General Surgery-VUH**
10/7/2013 to 10/13/2013

Evaluator
**Jonathan Wanderer**
Faculty

*Evaluate resident performance in each of the general competencies based upon (1) progress toward competent, unsupervised practic of anesthesiology and perioperative medicine and (2) performance relative to the resident's peer group. Also, indicate and comment about any substantial concerns warranting prompt review by the Program Director and Clinical Competence Committee.*

### MEDICAL KNOWLEDGE

Possesses appropriate basic science, medical and clinical knowledge relevant to anesthesiology and perioperative medicine. Recognizes and corrects gaps in knowledge and expertise. Critically evaluates and applies medical knowledge to patient care.

| Needed Significant Assistance | Needed Moderate Assistance | Needed Minimal Assistance | Needed Very Infrequent Assistance | Performed in a Competent, Independent Manner | Proficient; Consultant to Other Physicians | Expert; Resource for Other Anesthesiologists | N/A |
|---|---|---|---|---|---|---|---|
| O | O | O | O | O | O | O | O |

Medical knowledge compared to peers.

| Distinctly Below Peer Level | Somewhat Below Peer Level | At Peer Level | Somewhat Above Peer Level | Distinctly Above Peer Level | N/A |
|---|---|---|---|---|---|
| O | O | O | O | O | O |

Comments

Remaining Characters: 5,000

### PATIENT CARE

**PERIOPERATIVE ANESTHESIA MANAGEMENT:** Provides safe, effective, timely and efficient management of anesthesia and perioperative care, including the following: preoperative assessment and preparation of patients; appropriate use of monitors; safe airway management; evidence-based anesthesia management; anticipation and management of perioperative problems.
**REGIONAL ANESTHESIA AND PAIN MANAGEMENT:** Demonstrates appropriate knowledge, judgment and technical skills in managing regional anesthesia/analgesia. Assesses and effectively manages acute and chronic pain disorders.
**Critical Care:** Provides appropriate consultative support for critically ill patients, including thorough assessment and differential diagnosis; comprehensive management plan; effective collaboration with critical care team members; safe and efficient transfer of patients.
**Key Elements:** Vigilance, adaptability, careful and thorough practice; sound clinical judgment; anticipation and planning for foreseen and unforeseen events; timely and effective communication; and coordination of care among team members.

| Needed Significant Assistance | Needed Moderate Assistance | Needed Minimal Assistance | Needed Very Infrequent Assistance | Performed in a Competent, Independent Manner | Proficient; Consultant to Other Physicians | Expert; Resource for Other Anesthesiologists | N/A |
|---|---|---|---|---|---|---|---|
| O | O | O | O | O | O | O | O |

## Appendix 1. (Continued)

Patient Care compared to peers.

| Distinctly Below Peer Level | Somewhat Below Peer Level | At Peer Level | Somewhat Above Peer Level | Distinctly Above Peer Level | N/A |
|---|---|---|---|---|---|
| O | O | O | O | O | O |

Comments

Remaining Characters: 5,000

TECHNICAL SKILLS: Demonstrates appropriate knowledge, judgment and technical skill in selecting and performing procedures indicated for safe, effective and efficient anesthesia and perioperative management.

| Needed Significant Assistance | Needed Moderate Assistance | Needed Minimal Assistance | Needed Very Infrequent Assistance | Performed in a Competent, Independent Manner | Proficient; Consultant to Other Physicians | Expert; Resource for Other Anesthesiologists | N/A |
|---|---|---|---|---|---|---|---|
| O | O | O | O | O | O | O | O |

Technical skills compared to peers.

| Distinctly Below Peer Level | Somewhat Below Peer Level | At Peer Level | Somewhat Above Peer Level | Distinctly Above Peer Level | N/A |
|---|---|---|---|---|---|
| O | O | O | O | O | O |

Comments

Remaining Characters: 5,000

**PRACTICE-BASED LEARNING AND IMPROVEMENT**

Reviews and reflects upon personal performance and patient outcomes in order to improve performance and patient care; recognize and corrects gaps in knowledge and expertise; locates, appraises and applies scientific evidence to improve patient care; seeks guidance and assistance when needed.

| Needed Significant Assistance | Needed Moderate Assistance | Needed Minimal Assistance | Needed Very Infrequent Assistance | Performed in a Competent, Independent Manner | Proficient; Consultant to Other Physicians | Expert; Resource for Other Anesthesiologists | N/A |
|---|---|---|---|---|---|---|---|
| O | O | O | O | O | O | O | O |

Practice-based learning and improvement compared to peers.

| Distinctly Below Peer Level | Somewhat Below Peer Level | At Peer Level | Somewhat Above Peer Level | Distinctly Above Peer Level | N/A |
|---|---|---|---|---|---|
| O | O | O | O | O | O |

Comments

Remaining Characters: 5,000

## Appendix 1. (Continued)

### PROFESSIONALISM

Demonstrates the following: ethical and moral behavior; respect for patients, families and co-workers; commitment to fulfilling professional responsibilities; commitment to excellence and professional development. Characteristics include honesty, integrity, respect, compassion, reliability, diligence and responsibility.

| Needed Significant Assistance | Needed Moderate Assistance | Needed Minimal Assistance | Needed Very Infrequent Assistance | Performed in a Competent, Independent Manner | Proficient; Consultant to Other Physicians | Expert; Resource for Other Anesthesiologists | N/A |
|---|---|---|---|---|---|---|---|
| O | O | O | O | O | O | O | O |

Professional behavior compared to peers.

| Distinctly Below Peer Level | Somewhat Below Peer Level | At Peer Level | Somewhat Above Peer Level | Distinctly Above Peer Level | N/A |
|---|---|---|---|---|---|
| O | O | O | O | O | O |

Comments

Remaining Characters: 5,000

### INTERPERSONAL AND COMMUNICATION SKILLS

Develops sound therapeutic relationships with patients and families; respects and considers diversity (culture, language, religion, gender, age) in interpersonal interactions. Employs effective listening, verbal, non-verbal and written communication skills; collaborat effectively with healthcare team members; maintains clear and concise medical records. Projects appropriate competence and confidence; maintains composure in stressful situations.

| Needed Significant Assistance | Needed Moderate Assistance | Needed Minimal Assistance | Needed Very Infrequent Assistance | Performed in a Competent, Independent Manner | Proficient; Consultant to Other Physicians | Expert; Resource for Other Anesthesiologists | N/A |
|---|---|---|---|---|---|---|---|
| O | O | O | O | O | O | O | O |

Interpersonal and communication skills compared to peers.

| Distinctly Below Peer Level | Somewhat Below Peer Level | At Peer Level | Somewhat Above Peer Level | Distinctly Above Peer Level | N/A |
|---|---|---|---|---|---|
| O | O | O | O | O | O |

Comments

Remaining Characters: 5,000

### SYSTEMS-BASED PRACTICE

Demonstrates knowledge of perioperative and related healthcare systems, works efficiently, and utilizes system resources to provide cost-effective patient care. Complies with system-wide policies designed to improve outcomes , e.g., Joint Commission Regulations, HIPPA and infection control procedures. Contributes to quality improvement activities.

## Appendix 1. (Continued)

| Needed Significant Assistance | Needed Moderate Assistance | Needed Minimal Assistance | Needed Very Infrequent Assistance | Performed in a Competent, Independent Manner | Proficient; Consultant to Other Physicians | Expert; Resource for Other Anesthesiologists | N/A |
|---|---|---|---|---|---|---|---|
| O | O | O | O | O | O | O | O |

Systems-based practice compared to peers.

| Distinctly Below Peer Level | Somewhat Below Peer Level | At Peer Level | Somewhat Above Peer Level | Distinctly Above Peer Level | N/A |
|---|---|---|---|---|---|
| O | O | O | O | O | O |

Comments

Remaining Characters: 5,000

**PATIENT CARE**

What are the particular strengths of this resident?

Comments

Remaining Characters: 5,000

Residents want constructive advice for improvement. Please note specific areas or clinical skills for this resident to focus on for improvement:

Comments

Remaining Characters: 5,000

**OVERALL PERFORMANCE**

*Please document for the Program Directors and Clinical Competence Committee any serious concerns about this resident's attitudes, behaviors or performance.*

When supervising this resident, do you have concerns about patient safety?

| Yes | No | N/A |
|---|---|---|
| O | O | O |

Do you have concerns about this resident's honesty, integrity, reliability or character?

| Yes | No | N/A |
|---|---|---|
| O | O | O |

Are you concerned that this resident lags behind peers and needs extra help?

| Yes | No | N/A |
|---|---|---|
| O | O | O |

## Appendix 1. (Continued)

Are you concerned about this resident's teamwork and/or communication skills?

Yes ○ No ○ N/A ○

Are you concerned about this resident's openness to teaching and improvement?

Yes ○ No ○ N/A ○

Comments

Remaining Characters: 5,000

Have you discussed your concerns with this resident?

Yes ○ No ○ N/A ○

*Do you have confidence that this resident can competently manage the following cases without supervision?*

MAC for skin biopsy (healthy patient)

Yes ○ No ○ N/A ○

General anesthesia with an LMA for knee arthroscopy (healthy patient)

Yes ○ No ○ N/A ○

Appendectomy (full stomach, otherwise healthy)

Yes ○ No ○ N/A ○

Exploratory laparotomy for perforated viscus (otherwise healthy patient)

Yes ○ No ○ N/A ○

Elective colectomy in patient with coronary artery disease and renal insufficiency

Yes ○ No ○ N/A ○

Elective craniotomy in patient with increased intracranial pressure and asthma

Yes ○ No ○ N/A ○

Blunt trauma including head injury and liver injury

Yes ○ No ○ N/A ○

## Appendix 1. (Continued)

Leaking abdominal aneurysm in patient with decompensated congestive heart failure and atrial fibrillation

Yes ○ No ○ N/A ○

Confidential Comments:(will only be seen by the subject's Program Director)

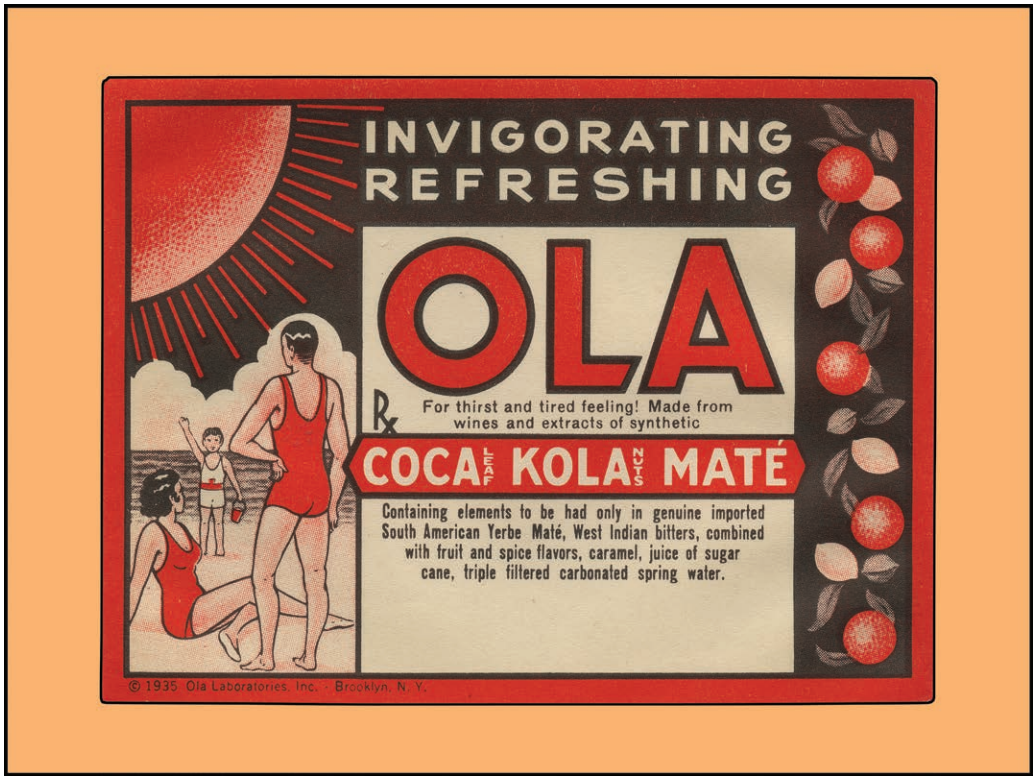
Remaining Characters: 5,000

Check this box to certify that you are **Jonathan Wanderer** and confirm your digital signature for this document on 10/15/2013.

Submit Final | Save Draft | Save Draft and Print | Submit as NET (Not Enough Time)

## ANESTHESIOLOGY REFLECTIONS FROM THE WOOD LIBRARY-MUSEUM

# A Cocaine Beverage…from Brooklyn: Ola Laboratories' Spicy Blend of Cola, Coca, and Maté




Even though the United States government had tightened restrictions on the public's access to cocaine and coca leaf products, Brooklyn's Ola Laboratories, Inc., copyrighted in 1935 its "invigorating" drink as "Ola" (*above*). The beverage blended "coca leaf, kola nuts, [and yerba] maté," flavored with fruit, spices, and bitters and combined with caramel, sugar cane juice, and carbonated water. By combining one of North Americans' most popular beverage combinations—cola with coca— with one of South Americans' favorites—yerba maté, Ola should have been a carbonated sales sensation. Instead its rollout fizzled, and the Brooklyn firm was absorbed by Len-Ola Laboratories. (Copyright © the American Society of Anesthesiologists' Wood Library-Museum of Anesthesiology.)

*George S. Bause, M.D., M.P.H., Honorary Curator and Laureate of the History of Anesthesia, Wood Library-Museum of Anesthesiology, Schaumburg, Illinois, and Clinical Associate Professor, Case Western Reserve University, Cleveland, Ohio. UJYC@aol.com.*