Consistency, Inter-rater Reliability, and Validity of 441 Consecutive Mock Oral Examinations in Anesthesiology

Implications for Use as a Tool for Assessment of Residents

Amin Schubert, M.D.,*† John E. Tetzlaff, M.D.,†† Ming Tan, Ph.D.,§ J. Victor Ryckman, M.D., † Edward Mascha, M.S.||

Background: Oral practice examinations (OPEs) are used extensively in many anesthesiology programs for various reasons, including assessment of clinical judgment. Yet oral examinations have been criticized for their subjectivity. The authors studied the reliability, consistency, and validity of their OPE program to determine if it was a useful assessment tool.

Methods: From 1989 through 1993, we prospectively studied 441 OPEs given to 190 residents. The examination format closely approximated that used by the American Board of Anesthesiology. Pass-fail grade and an overall numerical score were the OPE results of interest. Internal consistency and inter-rater reliability were determined using agreement measures. To assess their validity in describing competence, OPE results were correlated with in-training examination results and faculty evaluations. Furthermore, we analyzed the relationship of OPE with implicit indicators of resident preparation such as length of training.

Results: The internal consistency coefficient for the overall numerical score was 0.82, indicating good correlation among component scores. The interexaminer agreement was 0.68, indicating moderate or good agreement beyond that expected by chance. The actual agreement among examiners on pass-fail was 84%. Correlation of overall numerical score with in-training examination scores and faculty evaluations was moderate (r = 0.47 and 0.41, respectively; P < 0.01). OPE results were significantly (P < 0.01) associated with training duration, previous OPE experience, trainee preparedness, and trainee anxiety.

Conclusion: Our results show the substantial internal consistency and reliability of OPE results at a single institution. The positive correlation of OPE scores with in-training examination scores, faculty evaluations, and other indicators of preparation suggest that OPEs are a reasonably valid tool for assessment of resident performance. (Key words: Clinical judgment; competence; education; residency; testing.)

THE American Board of Anesthesiology (ABA) requires training programs to report on the performance of residents semiannually. Residency programs have few objective assessments of clinical competence at their disposal. General assessments by faculty evaluations can be limited by incomplete recall and personal bias, whereas written examinations primarily probe factual knowledge. Because oral examinations yield interactively obtained trainee responses to a fairly standardized, real-life clinical case scenario, they add to other evaluations by affording insight into trainees clinical reasoning, problem-solving ability, communication skills, and clinical judgment. In addition to providing a practice oppo-
tunity for the ABA oral examination process, oral practice examinations (OPEs) may therefore also represent a useful tool for assessment of clinical performance and learning progress.

However, oral examinations have been criticized for lack of reliability, validity, and objectivity. Specific areas of weakness are interexaminer agreement, case specificity (and, therefore, poor generalizability), undue influences by extraneous factors, and poor correlation with "objective" measures of knowledge. Fortunately, oral examinations in anesthesiology have been well refined and standardized under the aegis of the ABA in the last three decades. The format, strengths, and weaknesses of oral qualifying examinations in anesthesiology have been well communicated. Still, despite the widespread use of oral certification and mock oral examinations in the specialty, an evaluation of their internal consistency and reliability has not been reported since the original reports by Carter in 1962, and Kelley et al in 1971, when the ABA oral examination differed substantially from the current format. On the other hand, strong evidence was recently offered supporting the validity of the certification process, which, in addition to the oral examination, includes training requirements and written examination.

Given the widespread prevalence of OPEs and the substantial resources they require, we sought to examine if OPEs can also yield reliable and valid information useful to anesthesia educators who must evaluate resident performance. Accordingly, we determined OPE internal consistency and inter-rater reliability. To gauge validity with respect to competence assessment, we compared OPE scores with written in-training examination results and faculty evaluations. Furthermore, we analyzed the relationship of OPE with implicit indicators of resident preparation such as length of training and self-assessed preparedness. With the start-up of a comprehensive OPE program at our institution in 1989, we had the opportunity to address these issues prospectively in a well-controlled environment.

Methods

From spring 1989 to fall 1993, a formal semianual mock oral program, the OPE, was conducted using a set of 21 standardized question scenarios and a cadre of 17 faculty examiners. Although experience level varied among examiners, all were members of the professional staff with a permanent faculty appointment, had a demonstrated interest in resident education, had been board certified for at least 1 yr, and had attended at least one OPE inservice session yearly. Anesthesiology residents with at least 9 months of clinical anesthesia (CA) training participated in this mandatory activity. They received at least yearly training sessions on OPE format and on taking oral examinations. During the immediate preexamination period, candidates for the OPE completed a nonanonymous questionnaire assessing their level of anxiety, preparedness, and study habits. They completed an anonymous exit questionnaire immediately after the examination ended. During the study period, OPE performance was not used in the evaluation of any resident by either the program director or the clinical competence committee. A detailed description of the oral examination program at our institution is available elsewhere.

This program was modeled after the ABA oral examination. Therefore, it uses the guided-question format and includes a stem or case scenario that is divided into sections for preoperative evaluation, intraoperative management, and postoperative care. In addition to the stem, each question also contains an "additional topics" section containing two to three clinical vignettes designed to explore expertise in areas that are different from the material covered in the stem. Each examination session was preceded by a 7- to 10-min preparation period during which the candidate reviewed the short narrative of the stem question. The candidate was then asked to enter a faculty office and was seated with two faculty examiners. Examiner 1 began with the preoperative evaluation and questioned the candidate for 5 min. Examiner 2 continued for 15 min with the intraoperative and postoperative sections, followed by examiner 1, who explored one to three additional topics in depth for the next 10 min. This examiner returned to a previous question only if he or she felt that conduct of the examination did not allow a conclusive final grade to be assigned. At the conclusion of the examination, the resident was briefly excused and the examiners independently completed the standardized grading sheet. Thereafter, residents returned to the examination location for a 5- to 10-min debriefing session with the examiners.

A standardized grading sheet was used to guide examiner scoring. Both examiners independently rated the examinee's performance on the preoperative (A), intraoperative (B), postoperative (C), and additional topics (D) sections of the OPE. Each section (A-D) contained three to six subquestions that were graded separately. Examiners were told to score only the subquestions that
were discussed during the examination. A numerical "subscore" score was generated for each subquestion. Subscores were combined to yield the section score. Permissible subscores were 1, 2, 3, and 4, with 1 being the highest and 4 being the lowest.

Each examiner assigned a final grade to the candidate using a four-point scale (80, 77, 73, or 70). A grade of 80 was defined as a definite pass, and 70 was defined as a definite fail. In arriving at the final grade, examiners gave more weight to subquestions that they dealt with in greater detail. Examiners were urged to accumulate sufficient information during the examination to be able to distinguish clearly between pass and fail. Only in situations where the examining procedure left room for uncertainty were the grades of 77 and 73 to be used. In their written instructions, examiners were urged to refrain from questions that demanded exclusively factual information because such questions were considered ungradable. The stated objective of examiner questioning was to elicit the candidate's thought processes and evidence for a consultant level of function.

The OPE scoring procedure provided the inputs for the three major indices of candidate performance: (1) the pass-fail grade (a pass was assigned when the average of the two examiners' final grades exceeded 75); (2) the section score (calculated for each of sections A-D as the sum of all reported subscores in that section divided by the number of subquestions); and (3) the overall numerical score (ONS; defined as the sum of all subscores divided by the number of subquestions). It should be noted that computation of ONS gives each recorded subquestion score an equal weight, which is not the case in the examiners' choosing the pass-fail grade. In contrast to our OPE, numerical subscores or ONS are not reported in the ABA oral examination.

To begin to describe the overall reliability of the OPE, its internal consistency (the relationship among test scores on different parts of the test) was assessed. Internal consistency of the OPE was calculated using a generalized version of Cronbach \( \alpha \) to account for within-subject correlation, such as may have resulted from the fact that residents generally took more than one OPE. Cronbach \( \alpha \) is a measure of internal consistency that indicates the tendency of candidates who do well or poorly in one OPE section to perform similarly in another. It is calculated as a function of the number of subscores in the examination (here four), the sum of the variances of the subscore, and the sum of the covariances (hence, correlation) among the subscores. In our generalized version, we adjusted for the examiner, the number of previous examinations the resident had taken, and the resident him/herself when calculating the variances and covariances. This index should be approximately 0.9 for a certifying examination used to distinguish acceptable from unacceptable candidates, although a lower value (0.7) may be considered adequate for less crucial uses. A value of zero would mean that the subscores are not correlated at all, and a value close to -1 would mean that, on average, the items have a high negative correlation with each other. To assess internal consistency further, correlation between scores for all four OPE sections and ONS were assessed using Spearman correlation and general estimating equation analysis. Similarly, the reliability of the subscores was assessed by calculating their correlation with final grade. The general estimating equation method was used to obtain a correlation coefficient that takes into account the inherent dependency among repeated observations from the same subject.

A further indication that an examination is reliable is found by looking at its inter-rater reliability (the agreement between scores independently assigned by the two concurrent examiners). We calculated a generalized version of the intraclass correlation coefficient ranging from 0 to 1 that summarizes the proportion of the total variance explained by factors other than the examiners, such as the resident and number of previous examinations the resident had taken. This coefficient is similar to the \( \kappa \) statistic for independent observations. Values close to 1.0 indicate that in comparison to the other factors, the examiners scores on a particular examination do not vary much. We used a nested random effects model because we wanted to assess the variability among scores by the two examiners at each examination ("nested" within examination) and because both the examiners and residents should be considered as "random effects" (i.e., we were interested in generalizing to all residents and examiners, not just the ones in the study). In addition, a variation of the method of Bland and Altman, called the clinical agreement rate, was used to describe more specifically agreement in ONS between raters. The clinical agreement rate was defined as the percent of ONS examiner pairs that were within 0.75 points of each other, a cut point of clinical relevance, which we defined in advance. A clinical agreement of > 80% was considered good, meaning that the great majority of examiners differed by no more than 0.75 ONS points on a scale of 1 to 4.

Validity was examined, in part, by computing correlation coefficients between OPE outcomes and in-training
examination (ITE) results. ITE scores were the primary index of resident performance against which OPE results were compared because no other measure of residency performance has been validated similarly. Spearman rank-order correlation coefficients were computed between OPE grade and ITE scores for CA levels 1 to 3. This was performed initially for each examiner’s final score as well as when both examiners’ scores were combined into the final grade. However, inspection of interim results showed that there was no material loss of information when scores from both examiners were combined. Inter-rater comparisons combined with internal consistency data suggested that pass-fail grade and ONS are the most reliable and representative of OPE outcome scores. Therefore, ONS and pass-fail grade were used as surrogate OPE outcomes for all analyses.

In-training examination scores provide only a limited assessment of competence for anesthesiology residents. This prompted us to augment our assessment of OPE validity by referencing it to faculty evaluations. Toward the end of the 1991-1992 and 1992-1993 academic years, all anesthesia faculty from the operating room were asked to rate the clinical performance of individual residents. Faculty were asked to evaluate residents by cohort groups defined by their year of training. Within each cohort, faculty selected as anchors the best and worst resident and assigned them the scores of 100 and 0, respectively. They then rated all others in the cohort in relation to these anchors. All faculty rankings were then combined and averaged for each trainee whose OPE occurred during the ranking period. This ranking score was then assessed for association with ONS and pass-fail grade.

The following indicators of resident preparation were chosen as candidates for the analysis assessing association with the ONS and pass-fail OPE outcomes: length of training, prior OPE experience, self-assessed preparedness and anxiety, and prior exposure to either clinical rotations or literature that was related to the examination material. To account properly for training of transfer residents and for residents who started their training midyear, we used both CA level and the more accurate “months-in-training” to describe training duration. Months-in-training was defined as the total number of months the resident had been enrolled in an anesthesiology training continuum since the start of his or her CA-1 year, measured at the time of the OPE date and rounded to the nearest month. To determine the relationship of clinical exposure to OPE results, two faculty observers independently rated each OPE case scenario as either related or unrelated to the educational content (e.g., neuroanesthesia, cardiac anesthesia, and so on) of clinical rotations taken by the examinee during the 3 months before the OPE. Disagreements were resolved by a third faculty member. The section scores for parts A-D of the examination questions and ONS for examinees with exposure to a related recent clinical rotation were then compared with that of those who had unrelated clinical rotations. To investigate the influence of formal didactic experiences on OPE performance, each sub-question within the 21 OPE case scenarios used during the study period was rated by two independent faculty observers with respect to its content relationship to all lectures given to residents within 1 month before the OPE date. Only residents who regularly attended the didactic lecture series were included in this analysis. The association of pass-fail grade and recent OPE-related clinical experience was tested by general estimating equation methodology for the 338 examinations in 113 CA-1 and CA-2 residents. Self-assessed anxiety and preparedness scores were created by averaging the responses to applicable questions from the pre-OPE questionnaire.

To assess further the relationship of different indices of resident performance to OPE results (multivariable analysis), the general estimating equation method was again used for the pass-fail outcome, whereas a mixed-effects analysis of variance model was used for the ONS outcome. Thus, through the use of statistical models for ONS and pass-fail grade, we characterized the dependence of these OPE scores on the set of covariables comprising the indices of resident preparation explained in the preceding paragraph. As in the reliability analyses, these methods also account for the within-subject correlation attributable to a resident who took multiple OPEs. For the ONS model, the adjusted means (± SD) for binary covariables (yes/no) are reported. Furthermore, a correlation statistic that adjusts for correlation within subjects is provided to describe the relation between ONS and ordinarily scaled covariables. The ONS variance explained by the variable is the square of the adjusted correlation reported. The relation of pass-fail and the covariables of interest are given as odds ratios (and 95% confidence intervals) for each independent variable. The final statistical models for ONS and pass-fail grade included covariables that maintained a statistical significance of $P < 0.05$ in the model, while adjusting for other significant covariables. The extent to which the pass-fail model fit the data was checked using residual plots with simulated envelope, the absolute values of...
Results

During the period from April 1989 through October 1993, a total of 441 OPEs were administered to 190 residents. Of these, 116 took the OPE twice, 72 three times, and 63 four or more times (table 1). A total of 251 of 441 OPEs (57%) were taken by residents who had been examined previously. By the end of the study period, all except two OPE questions had been applied to ≥10 candidates.

The generalized Cronbach α was 0.82 for the four section scores on which the ONS is based, indicating good internal consistency. Section scores correlated moderately well with ONS or final grade, the Spearman correlation coefficients with final grade ranging from −0.38 to −0.67 (P < 0.01 for all section scores). Generalized inter-rater reliability coefficients for final grade, ONS, and pass-fail grades were 0.72, 0.65, and 0.68, respectively. For example, a final grade reliability coefficient of 0.72 means that only 28% of the variability of the observed grade is explained by interexaminer variability. For pass-fail grade, the observed agreement between two examiners was 84%; for ONS, the clinical agreement rate was 82%.

Examinees who passed the OPE had higher ITE scaled scores than those who failed (32.7 ± 6.7 vs. 27.6 ± 5.8, respectively; P < 0.01). ONS (r = −0.47) and pass-fail grade (r = 0.46) were also significantly correlated with ITE (P < 0.01). Note that the negative correlation values are artificially created by the scoring system used for subsections. A total of 43 faculty evaluated 145 residents and fellows. Eighty faculty ratings of residents could be matched with OPE results obtained concurrently during the faculty rating period. ONS (r = −0.43) and pass-fail grade (r = 0.62) were correlated with faculty ratings (P < 0.01). For comparison, ITE performance correlated with faculty rating at a level of r = 0.38 (P < 0.01). Faculty ratings of residents who passed the OPE were higher than for those who failed (69.4 ± 19.1 vs. 56.5 ± 20.6; P < 0.01).

Table 1 shows the relationship between OPE pass-fail score and the number of prior examinations taken. The chance of passing a repeat OPE was significantly higher (147 of 251 [58%]) than passing the first OPE (74 of 190 [39%]; P < 0.01). On average, those with previous OPE experience had ONS scores 0.24 points better (lower) than those without such experience, even after adjusting for anxiety and preparedness. In addition, the correlation among the pass-fail outcomes of repeated OPEs was 0.44 (P < 0.001), which lends support to our strategy of taking the correlation among repeated examinations of the same resident into account in all regression models. Regression analysis also showed a highly statistically significant univariate association between CA level and OPE pass rate (P < 0.001). OPE pass rates for CA-1, CA-2, CA-3, and CA-4 residents were 22 of 66 (33%), 73 of 135 (54%), 93 of 180 (52%), and 33 of 60 (55%), respectively. Residents who passed the OPE had spent more time in anesthesiology residency training (26 ± 11 months) when compared with those who failed (24 ± 11 months; P = 0.01). The number of months spent in anesthesiology training was also significantly correlated with ITE scores (r = 0.50; P < 0.01).

Residents' exposure to clinical rotations or didactic presentations with subject matter related to OPE ques-

Table 1. Pass–Fail Rate (%) as a Function of the Number of OPEs Taken

<table>
<thead>
<tr>
<th>Frequency of OPEs Taken</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fail</td>
<td>116</td>
<td>53</td>
<td>30</td>
<td>17</td>
<td>3</td>
<td>1</td>
<td>220</td>
</tr>
<tr>
<td>Pass</td>
<td>74</td>
<td>63</td>
<td>42</td>
<td>24</td>
<td>11</td>
<td>7</td>
<td>221</td>
</tr>
<tr>
<td>Total</td>
<td>190</td>
<td>116</td>
<td>72</td>
<td>41</td>
<td>14</td>
<td>8</td>
<td>441</td>
</tr>
</tbody>
</table>

* Tetrachoric correlation = 0.44 (P < 0.001).
Mock Oral Examination Validity and Reliability

Table 2. Relationship of Pass-Fail Grade and Attributes Indicative of Trainee Preparation

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Obs/Res</th>
<th>Univariate Analysis</th>
<th>Multivariable Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Odds Ratio*</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(95% CI)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>P Value</td>
<td></td>
</tr>
<tr>
<td>Months in training</td>
<td>441/190</td>
<td>1.02 (1.00, 1.04)</td>
<td>0.022 NS</td>
</tr>
<tr>
<td>Previous OPE (yes/no)</td>
<td>441/190</td>
<td>2.0 (1.39, 2.90)</td>
<td>&lt;0.001 NS</td>
</tr>
<tr>
<td>Prior related rotations (yes/no)</td>
<td>441/190</td>
<td>1.16 (0.81, 1.65)</td>
<td>NS NS</td>
</tr>
<tr>
<td>Prior related lectures (yes/no)</td>
<td>201/103</td>
<td>0.52 (0.30, 0.92)</td>
<td>0.025 NS</td>
</tr>
<tr>
<td>Anxiety†</td>
<td>353/169</td>
<td>0.83 (0.74, 0.93)</td>
<td>0.002 0.018</td>
</tr>
<tr>
<td>Preparedness†</td>
<td>342/165</td>
<td>1.34 (1.18, 1.52)</td>
<td>&lt;0.001 &lt;0.001</td>
</tr>
<tr>
<td>Case-related reading†</td>
<td>408/178</td>
<td>1.12 (0.91, 1.39)</td>
<td>NS NS</td>
</tr>
</tbody>
</table>

CI = confidence interval; Obs = observations; Res = residents; NS = not significant.
* Odds of passing OPE.
† Self-assessed; from questionnaire.

Cl - confidence interval; Obs - observations; Res - residents; NS = not significant.

The OPE pass rate was somewhat lower for residents who had OPE subquestions that dealt with lecture material compared with the pass rate of residents who did not have related lectures (42% vs. 57%; P = 0.025), but this relationship was not important in the multiple regression model after adjusting for other factors. Greater self-reported preparedness and lesser anxiety level were significantly related to better OPE performance, expressed as pass-fail grade or ONS (tables 2 and 3). Given the observed odds ratios, one point on the preparedness or anxiety scales, therefore, was associated with estimated 38% and 18% higher chances of passing, respectively. Preparedness was also univariately related to ITE score (r = 0.43; P < 0.001) and months in training (r = 0.40; P < 0.001).

In summary, univariate analyses yielded the following factors associated with both pass-fail and ONS OPE outcome: months in training, previous OPE exposure, case-related lectures, and self-assessed anxiety and preparedness (tables 2 and 3). The multivariable approach yielded a significant association for previous OPE experience, anxiety, and preparedness.

Discussion

This is the first study on systematically conducted anesthesiology OPEs. Although the calculated overall numerical OPE score is not used in the ABA examination, it seemed to represent an acceptable alternative to, or surrogate for, the pass-fail score because most of our observations held true for both measures. Our results indicate that OPEs can have very acceptable internal consistency and inter-rater reliability. Furthermore, OPE scores correlated moderately well with both faculty evaluations and in-training examination ITE scores and were related to anesthesia training duration, exposure to prior

Table 3. Relationship between ONS and Attributes Indicative of Trainee Preparation

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Sample Obs/Residents</th>
<th>Univariate</th>
<th>Multivariable</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Correlation*</td>
<td>Adjusted Means ± SE</td>
</tr>
<tr>
<td>Months in training</td>
<td>441/190</td>
<td>R = -0.14</td>
<td>N/A</td>
</tr>
<tr>
<td>Previous OPE (yes/no)</td>
<td>441/190</td>
<td>N/A</td>
<td>yes = 2.07 ± 0.05</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>no = 2.32 ± 0.05</td>
</tr>
<tr>
<td>Prior related rotations (yes/no)</td>
<td>353/190</td>
<td>N/A</td>
<td>yes = 2.16 ± 0.05</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>no = 2.23 ± 0.05</td>
</tr>
<tr>
<td>Prior related lectures (yes/no)</td>
<td>201/103</td>
<td>N/A</td>
<td>yes = 2.41 ± 0.06</td>
</tr>
<tr>
<td>Anxiety</td>
<td>353/169</td>
<td>R = 0.16</td>
<td>N/A</td>
</tr>
<tr>
<td>Preparedness</td>
<td>342/165</td>
<td>R = -0.34</td>
<td>N/A</td>
</tr>
<tr>
<td>Case-related reading</td>
<td>408/178</td>
<td>R = -0.08</td>
<td>N/A</td>
</tr>
</tbody>
</table>

N/A = not applicable; NS = not significant.
* Analogous to a Pearson correlation coefficient, but adjusted for within-subject correlation.

Anesthesiology. V 91. No 1, Jul 1999
OPEs, and examinee-assessed preparedness. These observations suggest that OPEs provide a reasonable indication of residents’ acquired clinical competence.

Reliability of an examination is a measure of the test's precision. Reliability also assesses how consistently the same result is observed under different circumstances. In this context, we analyzed both internal consistency (overall reproducibility) and inter-rater reliability (reproducibility across two examiners). Our indicator of internal consistency was 0.82, a level considered adequate for most purposes. In addition, successful oral examinations more structured than our OPE still only show the degree of consistency observed in this study.

The high internal consistency of our OPE may indicate that examination questions were generally of uniform quality, eliciting similar responses in their component parts. Better students are thought to perform better than less-able students across a wide range of content and skills, which likely also occurred in our OPEs. An alternative explanation for the observed internal consistency would invoke a “halo” effect. It might be argued that candidates who were known to faculty examiners as good performers were given high OPE marks in all examination sections independently of true OPE performance. However, the acceptable agreement between live and taped ONS and pass-fail scores suggests that the halo effect did not influence our observations in a major way. Further evidence against a substantial halo effect lies in the observation that section scores and final grade were modestly but not highly correlated.

Assessment of inter-rater reliability is important because it indicates the reproducibility of the OPE grading method and can be considered one measure of examination quality by which OPE efforts are judged. Furthermore, it allows comparison with published reliability characteristics of other oral examinations. Inter-rater reliability may be improved by accounting for the grading patterns of individual examiners (“doves” vs. “hawks”). In our study, such a correction or normalization could only have been accomplished for the continuous variable ONS but was thought unnecessary. Rater characteristics have been found not to affect structured surgical oral examinations. Recognized means of improving the reliability of oral examinations include standard grading forms, examiner training, grading from videotaped recordings and standardized, scripted questions to reduce content sampling error, and the use of two simultaneous evaluators. Our OPE includes many of these features, perhaps accounting for the very acceptable observed reliability.

Oral examinations have been criticized because of unreliability stemming from the subjective nature of examiner scoring, the variability of performance based on the case scenario presented, and the influence of factors extraneous to the candidate’s competence. Our generalized and chance-corrected OPE inter-rater reliability ranged from 0.65 to 0.72. An interexaminer correlation of 0.7 may be acceptable depending on the situation. This level of reliability compares adequately with previously reported oral certifying examinations that are expected to achieve reliabilities close to 0.9. Thus, the inter-rater reliability was 0.80 for oral board examinations in emergency medicine and ranged from 0.63 to 0.83 for previous versions of the oral board examination in anesthesiology. Furthermore, our OPE examiners agreed on pass-fail grading decisions 84% of the time, compared with 67% in the certifying oral examination of the American Board of Psychiatry. Therefore, OPE inter-rater reliability was more than satisfactory for the purposes of an internal examination and compared favorably with previously reported inter-rater reliability of oral specialty board examinations.

A test is considered to have validity when it measures what it is intended to measure. The characteristic that the examination claims to measure is called a construct. Therefore, construct validity refers to the validity of a test with respect to a known, well-defined property, which is measured by a “gold standard” or, at least, by another credible measure. In addition to having construct validity, a test can also be considered valid in other ways. Face validity refers to an intuitive validity based on whether the test can reasonably be expected to measure the area of interest. Content validity holds when the test questions cover the area to be tested in a representative manner. Discriminant validity refers to

Anesthesiology, V 91, No 1, Jul 1999
the ability of a test to distinguish between novices and experts.

Our study assesses OPE validity in a number of ways. Considering clinical competence as the relevant construct, we assessed OPE results against written anesthesiology examinations, faculty evaluations, and supporting indicators of preparation. With the possible exception of faculty ratings, none of these measurements constitutes a gold standard, but they have been the subject of previous investigation and are operant in many anesthesiology training programs. For face and discriminant validity, we considered length of training, prior oral examination exposure, resident self-assessment, and, to a limited extent, the absence of confounding factors such as differing proximate educational experiences. Content validity of the OPE was addressed during the process of constructing the case scenarios. The authors of OPE case scenarios were veteran oral examiners and based the content of the examination questions on their ABA or OPE experience. Furthermore, directed efforts were made to achieve a diverse sampling of anesthetic problems within any one case scenario and through the use of additional topics.

Based on comparisons with ITE score and faculty evaluations, our OPE results can be said to show a moderate level of construct validity. The relationship between OPE outcomes and ITE scores was significant \( r = 0.47 \) and consistent with other reports from medical training programs. \(^{16,21,25,26} \) A similar level of correlation \( r = 0.42-0.54 \) has been reported between ABA written and oral examinations, suggesting that OPEs can approximate an ABA oral examination in this respect.

Our OPE results were moderately correlated with faculty evaluations, which continue to play a large role in resident evaluation. Subjective faculty assessment of residents has been used as a measure of clinical performance, \(^{26-29} \) despite some criticism. \(^{30,31} \) Colliver acknowledges that faculty ratings represent a proxy gold standard for clinical competence. ABA certification, which includes oral examinations, was recently validated on assessments by anesthesiology residency program directors, \(^7\) which constitutes a form of faculty evaluation. Our observation that faculty ratings were significantly correlated with ITE scores has been noted in previous studies of other global rating scales, \(^{26,52} \) further supporting the value of our faculty ratings. Nevertheless, the use of global faculty evaluations presents a number of limitations. In our dataset, exposure of residents to faculty raters may have been variable. Raters may rank individuals based on poor recollection or selective memory. Without anchors in the rating scale (as used in the present study), differences in raters' standards may invalidate the results. Skeptics further argue that subjective test performance and global faculty evaluations can be related because of a potential halo effect. However, with respect to the halo effect, it should be noted that the authors' faculty rankings were obtained from a group of attending physicians more than twice as large as the OPE examiner faculty. Furthermore, acceptable agreement was obtained between ratings by the original examiners and other examiners who rated the same examination performance on videotape without prior knowledge of the candidate. Consequently, we interpreted the moderate observed correlation of OPE scores with faculty evaluations as meaning that OPEs can assess a portion of the competency attributes gauged by faculty evaluations (fig. 1).

Although the correlations of OPE scores with ITE results and faculty evaluations are only modest and do not by themselves indicate validity, OPE validity is further supported by the association of OPE scores with training duration, CA level, and preparedness. OPE pass rate increased with anesthesia training duration and greater exposure to the OPE process. This is consistent with the notion that skills and experience level important for performing on the OPE are accumulated during the entire training period. Self-assessed preparedness was significantly correlated with months in training and ITE performance. Trainee estimates of their own performance previously have been found to be reasonably accurate in oral examinations for psychiatry residents. \(^{28} \)

Similarly, medical student feedback on neurology rotations has correlated well with objective structured clinical examination performance. \(^{33} \) However, self-assessment information should be interpreted with the knowledge that the trainee questionnaires were not completed anonymously. It is further acknowledged that other characteristics of test takers such as gender, eye contact, speaking time, and pace of delivery can affect grades. \(^{34-35} \)

More extensive previously gained clinical experience has been associated with better oral examination results and subjective evaluations of medical students on their surgery clerkship. \(^{37} \) We chose to measure recent clinical experiences by the examinees' participation in subspecialty block rotations. The substantial surgical caseload at this institution spawned anesthesia "sections" with corresponding resident rotations in almost every surgical subspecialty. Sections have their own educational structure, frequently including a syllabus, small group sem-

Anesthesiology, V 91, No 1, Jul 1999
Anesthesiology, V 91, No 1, Jul 1999

SCHUBERT ET AL.

Fig. 1. Schematic of a conceptual framework suggesting how available evaluation methods can combine to yield a comprehensive picture of trainee ability through assessment of various abilities comprising resident competence. The modest correlation \( r = 0.47 \) between OPE and written examination scores may indicate that the performance on the OPE presupposes a certain foundation of factual knowledge. This is schematically represented by the overlap between the ITE and OPE assessment areas. Faculty evaluation assesses interpersonal skills and technical competency but takes into account other skills and traits. This view is represented graphically by the overlap between the faculty, ONS, and ITE evaluation areas. It is supported by the modest but not insignificant observed correlation \( r = 0.43 \) between faculty ratings and OPE performance. The fact that performance scores representing each of the three evaluation methods are only modestly correlated likely means that each method evaluates a somewhat different skill set. However, it is unknown how much of gap between the observed correlation and a "perfect" correlation (e.g., \( r = 0.95-1.0 \)) is caused by variability unrelated to resident performance.

An evaluation tool should track performance and, in the case of anesthesiology residency training, should document performance improvement with accumulating educational experience. Our OPE results showed a positive relationship with a number of direct (ITE, faculty evaluations) and indirect (length of training, cumulative exposure to OPEs, and self-assessed preparedness) indicators of residents' preparation. This suggests that OPE performance can indicate residents' level of preparation for competent independent practice and board certification. Structured oral examinations have been found previously to predict clinical performance in pediatric house officers and candidates for American Board of Emergency Medicine certification. Therefore, anesthesiology educators may glean important information about their residents' preparedness from OPEs, which may spawn remediation or curricular enrichment in the areas of competence assessed by OPEs. Finally, the consistency and inter-rater reliability data from our OPE should help educators defend the use of OPEs as an evaluation tool with regard to trainees who may question oral examinations on the basis of high variability and arbitrariness.

Interpretation of our results must take into account several limitations. The information was derived from a single residency program with a well-organized oral practice examination effort. For programs in which sim-
ilar resources and procedures are not in place, our conclusions may not apply. Although comparison of OPE scores with ITE and faculty evaluations were made, no comparison with eventual board certification status was undertaken. Furthermore, the new format for the ABA oral examination adopted in 1997 puts more emphasis on perioperative evaluation and management. Therefore, if reliability and validity of the OPE depend critically on variations in format and content, our results may not be directly applicable in the new environment. To fulfill the most stringent requirements for reliability and validity, the OPE will likely require further development and standardization. The standard against which the examinees responses are graded may need better definition, and a greater number of case scenarios should be presented to each candidate to lessen the effect of content specificity on reliability. Although these improvements require careful planning and dedicated faculty support, they seem preferable to developing an entirely new system of evaluation given the existing prevalence of OPEs.

Our results add to the growing evidence that anesthesiology oral examinations in the ABA format can yield a valid indication of trainee performance. Resident assessment through case-oriented mock oral examinations can give anesthesia educators a tool to assess critical traits that are only partially gauged by other methods. Educators can be reassured that carefully conducted mock oral examinations are characterized by acceptable internal consistency, inter-rater reliability, and moderate construct and face validity.

The authors thank the following individuals without whom the OPE program would not have been possible: anesthesiology residents, fellows, and faculty of The Cleveland Clinic Foundation; Shelly Sords, Education Coordinator of the Division of Anesthesiology, for cheerful support and interest; Charlie Androjna for skillful management of the large database; Dr. Frances Rhoton for encouragement and support during the early days of the project; and Dr. Arthur Barnes, Vice Chairman, and Dr. Fawzy G. Estafanous, Chairman of the Division of Anesthesiology, for creating an environment in which the OPE program could flourish. We also thank Loric Peterson, Ronnie Sanders, and Nikki Williams for expert secretarial support. We would like to recognize Dr. John Ammon, Vice President of the American Board of Anesthesiology, and Dr. Francis Hughes, Executive Vice President of the American Board of Anesthesiology, for reviewing the manuscript and providing important clarifications and suggestions. We are indebted to Drs. Gerald Burger, Dorothea Markakis, Brian Parker, Michael O’Connor, and Paul Potter for their review and independent scoring of OPE videotapes.

Anesthesiology, V 91, No 1, Jul 1999

References

37. Stilman RM: Effect of prior clinical experience on students' knowledge and performance in surgery. Surgery 1986; 100:77-82
41. Elstein AS: Beyond multiple-choice questions and essays: The need for a new way to assess clinical competence. Acad Med 1993; 68:244-9