General Anesthesia

The inter-rater and intra-rater reliability of a new Canadian oral examination format in anesthesia is fair to good

[La fiabilité interexaminateurs et intra-examinateurs d'un nouveau modèle d'examen oral canadien en anesthésie est de moyenne à bonne]

Ramona A. Kearney MD FRCPC,* Stephen A. Puchalski MD FRCPC,† Homer Y.H. Yang MD FRCPC,† Ernest N. Skakun PhD*

Purpose: In response to the Royal College's request to improve the validity and reliability of oral examinations, the Examination Board in anesthesia proposed a structured oral examination format. Prior to its introduction, we studied this format in two residency programs to determine reliability of the examiners.

Methods: Twenty faculty and 26 residents from two Canadian residency programs participated (Sites A and B). Pairs of examiners scored five or six residents examined consecutively on two standardized questions using global rating scales with anchored performance criteria. Residents' performances were scored independently during the examination (Time 1) and later from a videotaped recording (Time 2). Correlations between scores of the pairs of examiners and between scores of each examiner were determined.

Results: Correlations demonstrating inter-rater agreement between examiners at Site A ranged from -.324 to .915 (mean .506) at Time 1. At Time 2, correlations ranged from .64 to .887 (mean .791). At Site B correlations ranged from .279 to .989 (mean .707) at Time 1 and at Time 2 correlations ranged from -.271 to .924 (mean .477).

Correlations demonstrating intra-rater agreement of examiners at Site A ranged from .054 to .983 (mean .723) and at Site B correlations ranged from -.055 to .974 (mean .662).

Correlations >0.4 were seen in 80% of the scores and >0.7 in 50% indicating fair to good intra-rater and inter- rater reliability using this format.

Conclusions: Despite the limitations of our study our results compare favourably with those previously reported in anesthesia. We recommend the adoption of this format to the Royal College of Physicians and Surgeons of Canada Examination Board.

Objectif : *C'est à la demande du Collège royal, d'améliorer la validité et la fiabilité des examens oraux, que le Bureau des examinateurs en anesthésie a proposé un modèle d'examen oral structuré. Avant sa mise en application, nous l'avons testé dans deux programmes de résidence afin de déterminer la fiabilité des examinateurs.*

Méthode : *Vingt facultés et 26 résidents de deux programmes canadiens ont participé à l'étude (Sites A et B). Des paires d'examinateurs ont utilisé une échelle de notation globale comportant des critères de rendement définis pour évaluer cinq ou six résidents appelés à répondre consécutivement à deux questions normalisées. Les résultats des résidents ont été cotés séparément pendant l'examen (Temps 1) puis, à partir d'un enregistrement vidéo (Temps 2). Les corrélations entre les scores des paires d'examinateurs et entre les scores de chaque examinateur ont été établies.*

Résultats : *Les corrélations démontrant une concordance interexaminateurs au Site A sont de -0,324 à 0,915 (moyenne de 0,506) au Temps 1. Au Temps 2, de 0,64 à 0,887 (moyenne de 0,791). Au Site B, elles sont de 0,279 à 0,989 (moyenne 0,707) au Temps 1, et au Temps 2 de -0,271 à 0,924 (moyenne de 0,477). Les corrélations sur la l'accord intra-examinateurs au Site A sont de 0,054 à 0,983 (moyenne de 0,723) et au Site B de -0,055 à 0,974 (moyenne de 0,662). Les corrélations étaient > 0,4 dans 80 % des scores et > 0,7 dans 50 %; la fiabilité intra-examinateurs et interexaminateurs ainsi*

From the Departments of Anesthesia, University of Alberta,* Edmonton, Alberta, and McMaster University,† The Division of Studies in Medical Education, Hamilton, Ontario, Canada.

Address correspondence to: Dr. Ramona A. Kearney, Department of Anesthesiology and Pain Medicine, University of Alberta, Clinical Sciences Building, Room 8-120G, Edmonton, Alberta T6G 2B7, Canada. Phone: 780-407-2689; Fax: 780-407-7461; E-mail: rkearney@ualberta.ca

Accepted for publication July 16, 2001.
Revision accepted November 16, 2001.

indiquée est de moyenne à bonne avec ce modèle.

Conclusion : *Malgré les limites de notre étude, nos résultats se comparent favorablement avec ceux qui ont déjà été signalés en anesthésie. Nous recommandons l'adoption du modèle par le Bureau des examinateurs du Collège royal des médecins et chirurgiens du Canada.*

THE Royal College of Physicians and Surgeons of Canada (RCPSC) has recommended that each specialty evaluate their summative assessments with the goal of providing supporting evidence for validity, reliability and efficiency.[a] The examination for certification from the RCPSC includes a written and oral component. Both components are designed to assess knowledge but while the written test is reliable and valid, it lacks realism as an examination of clinical competence. On the other hand the oral examination suffers from problems with reliability and objectivity which threaten its validity.[1] Potential sources of error include the items or cases chosen as well as the variability of the observer, in this case, the examiner.

While some specialties have rejected the oral examination in favour of other examination formats, the specialty of anesthesia has decided to maintain it believing that a structured oral examination can best evaluate the elements of problem solving. Many studies have shown low correlations between scores on oral and written examinations suggesting they may be measuring different aspects of competence.[1] The oral examination format at the time of this study consisted of two sessions in which three examiners asked the candidate five standardized questions over a one-hour period per session. All three examiners score the candidate on an anchored global rating scale, but only two of the examiners ask the questions and the third records the candidate's responses. The second session usually, but not always, has three different examiners. Despite independent scoring by the examiners, several common rating errors remain a concern. One of these is the halo effect where an examiner's overall judgement of a candidate is unduly influenced by one aspect of the candidate's performance such that the performance on one factor contaminates the judgement of performance on the other factors. The other is rater prejudice where there is tendency to rate positively certain types of candidates and negatively certain others.

[a] Royal College of Physicians and Surgeons of Canada. Newletter 1997; 7: 1.

A structured oral examination has advantages similar to that of an objective structured clinical examination (OSCE) in that the objectivity and reliability of the test is improved by minimizing patient variability with standardized patients or patient management problems, by using specified performance criteria and by minimizing examiner variability.[2] The purpose served by the present study was to investigate inter-rater and intra-rater agreement in this new oral examination format.

Methods

All residents in anesthesia from two Canadian residency programs who were required to sit the semi-annual departmental oral examination (those in PGY 2–5) were asked to participate. All agreed and provided written consent. Staff examiners were those who usually participated in departmental oral examinations and also provided written consent. In each program (Sites A and B), a structured practice oral examination took place. At site A, eight examiners and eight residents participated; at site B, 12 examiners and 18 residents participated. Examiners were faculty members from their respective universities and had experience in practice oral examinations in their programs. They were oriented to the scoring system and to the two questions asked in their session several days prior to the examination.

Examiners were paired for the examination, asked one question each and scored both questions. Questions were standardized, identical in format to those of the RCPSC examination and reviewed by two of the authors (RK, SP), current examiners for the Royal College. Scoring of the answers utilized the rating scale and performance criteria employed by the Royal College with the approval of the chief examiner. Examiners scored the answers independently and remained unaware of the scores of their co-examiners (Time 1). As this examination constituted the usual semi-annual departmental oral examination for the residents, verbal feedback on performance was given to each resident following scoring. Actual scores were not revealed to the residents.

Each pair of examiners scored six consecutive residents creating 12 scores each. All sessions were videotaped. Approximately two weeks later, examiners rescored the examinations (Time 2). Pearson product moment correlations between scores of the pairs of examiners were determined for each question for both Time 1 and Time 2. Scores of each individual examiner were compared from Time 1 to Time 2 and Pearson product moment correlations were also determined.

Immediately following the examinations, group interviews were held separately with the examiners and

TABLE I Inter-rater agreement on scores

Correlations based on six pairs of scores at Time 1 and Time 2						
Question	*Examiners*	*U of A*	*Examiners*	*McMaster*	*Examiners*	*McMaster*
Neuro	A/B	.822	E/F	.491	K/L	.974
		.823		.663		.701
Regional	A/B	.255	A/B	.279	G/H	.939†
		.698		-.259		.746†
Obstetrics	C/D	.733	A/B	.989	G/H	.898†
		.886		.924		.791†
Trauma	C/D	.915				
		.872				
Pediatric	E/F	.524	E/F	.354	K/L	.812
		.887		-.271		.882
Airway	E/F	.243	C/D	.645	I/J	.408†
		.834		.468*		.134†
Cardiac	G/H	.878	C/D	.801	I/J	.896†
		.64		.444		.505†
Respiratory	G/H	-.324				
		.687				

*1=session not videotaped at Time 2, comparisons of five sessions only. †1=resident no-show, comparison of five residents only.

TABLE II Intra-rater agreement on scores

Correlations based on six pairs of scores						
	University of Alberta		*McMaster University*		*McMaster University*	
	Examiner 1	*Examiner 2*	*Examiner 1*	*Examiner 2*	*Examiner 1*	*Examiner 2*
Neuro	.876	.702	.579	.548	.516	.836
Regional	.739	.193	.302	.346	.974*	.548*
Obstetrics	.815	.965	.788	.874	.938*	.877*
Trauma	.966	.803				
Pediatrics	.737	.933	.577	-.055	.873	.776
Airway	.833	.352	.612*	.845*	?	.913*
Cardiac	.73	.881	.71	.662	.313*	.456*
Respiratory	.054	.983				

*1=case deleted.

with the residents to obtain their opinions of the process. Specific questions to the examiners included: the effect of asking the same question repeatedly four to six times, the ease of adhering to a 25 min time limit, the use of standardized questions and the examiners' willingness to change to this format in the future. Questions to the residents included: the ease of understanding instructions in proceeding to various examination rooms, the effect of changing from a one hour examination of four to six questions to consecutive rooms with 25 min examinations of two questions, the use of standardized questions and their willingness to change to this format in the future.

Results

One resident at Site B did not present for the examination. One examiner at Site B, (examiner 1) did not score the videotapes at Time 2. In one session the examination was not videotaped.

Correlations between the pairs of examiners (i.e., examiners A and B ,etc.) are shown in Table I and summarized below:

For inter-rater agreement

Site A, Time 1: mean .506, median .626 (range -.324 to .915);
Site A, Time 2: mean .791, median .828 (range .64 to .887);
Site B, Time 1: mean .707, median .806 (range .279 to .989);
Site B, Time 2: mean .477, median .564 (range -.271 to .924).

Correlations between scores of individual examiners from Time 1 to Time 2 are shown in Table II. Examiner 1 is the examiner of the pair who is asking

the question and examiner 2 is the silent examiner for that question. The results are summarized below:

For intra-rater agreement
Site A: mean .723, median .809
(range .054 to .983);
Site B: mean .644, median .662
(range -.055 to .974).

Focused postexamination discussions with residents and staff revealed unanimous approval of the format and recommended adopting it. Supportive comments indicated that residents did not find the process confusing or disruptive and they were in favour of only two questions per examining team as this appeared to create a fairer examination. Staff examiners did not find the process induced fatigue or the inability to remain consistent from resident to resident. All found the quality of the standardized questions a marked improvement over non-standardized questions. There were no negative comments about the process.

Discussion

Summative assessments of knowledge are required by certification bodies to determine competency. Oral examinations in anesthesia assess the following previously described competencies: evaluating clinical situations, choosing therapies and justifying choices, dealing with changing situations, making decisions and communicating.[3] An editorial by Pope some years ago outlined perceived advantages of the oral examination format in anesthesia and felt the format should continue only if there were efforts by the Examining Board to "review, improve and educate itself… striving always for greater objectivity and…. validity.[4] The board has moved in this direction and as a result the examination has changed from Pope's day. Since then, all questions are standardized and global rating scales with anchored performance criteria are used. All scoring is independent and the final score is an average of all scores and not a pass/fail consensus in an attempt to minimize measurement error.

We were interested in determining the rater reliability of the examination by adopting the OSCE format. Improvement in rater agreement might be expected as, according to Muzzin and Hart, pairs of examiners have better reliability than individuals or larger groups and the longer the test, the more reliable it is.[1] Bias, in terms of the halo effect, should be significantly reduced when an examiner is scoring performance on two questions instead of five. The Examination Board was keen to have this format tested with respect to reliability and feasibility prior to its adoption by the Examination Board.

Reliability of the RCPSC oral examination in anesthesia has not been reported, so comparisons with the results of the present study cannot be made. Furthermore, although a study comparing the traditional format with the new format would be ideal to identify the magnitude of change in examination characteristics, the small number of residents in our programs would likely preclude finding significant differences even if they were present. Schubert reported inter-rater reliability of mock oral examinations in anesthesia on 441 oral practice examinations of 190 residents by 17 faculty examiners using a format similar to the American Board of Anesthesiology.[5] He reports inter- rater reliability as generalized reliability coefficients for final grade received and pass-fail grades as 0.72 and 0.68 respectively. Also reported is the inter-rater reliability on the overall numerical score (defined as the sum of all subscores divided by the number of subquestions) as 0.65. This compares favourably, as do our results, with previously reported results of American oral board examinations from three to four decades ago and with a review of the traditional oral examination literature.[6]

The use of this format in two residency programs as part of their oral practice examinations demonstrated inter- rater reliability with medians from .584 to .887 and similar levels of intra-rater agreement. Intra-rater agreement was lower at McMaster which may have been a factor in the decrease in inter-rater agreement seen there at Time 2. A wide range of correlations was seen including some negative correlations. There may be several reasons for this. The residents examined represented all levels of core anesthesia training with one quarter doing a practice examination in anesthesia for the first time. This could be expected to reduce examiner agreement. It has been shown that examiners are less consistent when rating poor performances.[7] As well, several authors allude to the importance of trained examiners.[3,8] Only one faculty member at each site was a current board examiner. The remainder varied in experience in practice examinations. The number of scores per examiner on which correlations were determined was quite small (five or six) which greatly influences the values obtained for reliability.[9] Sample sizes for most OSCE are much larger. In any event, the possibility of poor inter-rater correlations with certain questions emphasizes the need for a large number of questions for each candidate to reduce the likelihood of a poor decision in a high stakes examination.

The use of global rating scales may also be a factor in reduced examiner agreement. Checklists, by their

dichotomous nature, restrict the examiners' choices and would intuitively improve agreement. One must question however, whether higher order cognitive skills are amenable to assessment by checklists. Certainly, physician feedback regarding their experience with checklists has indicated concern that students could obtain high checklist scores without having appropriate approaches to problems.[10] In other words, the quality of the performance cannot be assessed with a checklist. Given that the competencies we wish to assess are present on a continuum in individual physicians, global rating scales appear most logical. Evidence is accumulating that when using physician examiners, global ratings are as reliable as checklists and have better construct and concurrent validity.[11] For assessing higher levels of competence, checklists have recently been challenged as valid measures.[12]

Positive aspects of this format were highlighted by both faculty examiners and residents. Most notable was the increase in perceived fairness by the residents. Since a poor performance on a question would not be known to the remaining examiners, the resident could remain confident of performing acceptably overall. Faculty examiners were aware of the necessity to present the question in the same way to consecutive residents which further standardized the process. In the previous format, examiners would ask the same question, at most, twice and often just once. The examiner is likely less aware of examiner fatigue or shifting standards in this way.

The examination was lengthened by this process due to transit time between station and more attention was required to organize it. The total number of examiners could likely remain the same. These findings were unlikely to have a significant impact on the structure and cost of the examination.

Determining the rater agreement for this examination looks at only one component of the reliability. There are other sources of variance in this format such as that between stations and examination candidates. Performing a generalizability study would help elucidate the relative importance of these factors and allow us to focus on which elements require attention. Future study of the nature of rater agreement should address some of the factors identified above such as examiner training which likely leads to improved agreement. In addition, standard setting for this examination should be scrutinized. We hypothesize that examiner agreement on the critical features of a candidate's answer which determines success or failure is crucial to obtain good rater reliability. This can be done through more careful construction of test items and a quality assurance review of items leading to consensus of critical features.

In conclusion, we evaluated a new format for the RCPSC oral examination in anesthesia and recommended its adoption to the Oral Examination Board. We also recommend ongoing assessment of the certification process to determine its psychometric characteristics and identify areas requiring improvement.

References

1 *Muzzin LJ, Hart L.* Oral examinations. *In*: Neufeld VR, Norman GR (Eds.). Assessing Clinical Competence. New York: Springer Publishing Co., 1985: 71–93.

2 *Harden RM, Gleeson FA.* Assessment of clinical competence using an objective structured clinical examination (OSCE). Med Educ 1979; 13: 41–54.

3 *Eagle CJ, Martineau R, Hamilton K.* The oral examination in anaesthetic resident evaluation. Can J Anaesth 1993; 40: 947–53.

4 *Pope WDB,* Anesthesia oral examination (Editorial). Can J Anaesth 1993; 40: 907–10.

5 *Schubert A, Tetzlaff JE, Tan M, Ryckman JV, Mascha E.* Consistency, inter-rater reliability, and validity of 441 consecutive mock oral examinations in anesthesiology. Implications for use as a tool for assessment of residents. Anesthesiology 1999; 91: 188–98.

6 *Anastakis DJ, Cohen R, Reznick RK.* The structured oral examination as a method for assessing surgical residents. Am J Surg 1991; 162: 67–70.

7 *Burchard KW, Rowland-Morin PA, Coe NPW, Garb JL.* A surgery oral examination: Interrater agreement and the influence of rater characteristics. Acad Med 1995; 70: 1044–6.

8 *Thomas CS, Mellsop G, Callender K, et al.* The oral examination: a study of academic and non-academic factors. Med Educ 1992; 27: 433–9.

9 *Streiner DL, Norman GR.* Health Measurement Scales. A Practice Guide to Their Development and Use, 2nd ed. Oxford: Oxford University Press, 1995: 114–6.

10 *Cohen R, Rothman AI, Poldre P, Ross J.* Validity and generalizability of global ratings in an objective structured clinical examination. Acad Med 1991; 66: 545–8.

11 *Regehr G, Freeman R, Robb A, Missiha N, Heisey R.* OSCE performance evaluations made by standardized patients: comparing checklist and global rating scores. Acad Med 1999; 74: S135–7.

12 *Hodges B, Regehr G, McNaughton N, Tiberius R, Hanson M.* OSCE checklists do not capture increasing levels of expertise. Acad Med 1999; 74: 1129–34.