Editorials

The oral examination process – gold standard or fool's gold

Patricia Houston MD MEd FRCPC,* Ramona A. Kearney MD MMEd FRCPC,† Georges Savoldelli MD MEd*

THE oral examination process has been a traditional component of the certification process in anesthesia for over 70 years.[1] It is believed that the oral examination can be used to evaluate problem solving ability, communication and collaboration skills and expert content knowledge skills. In spite of its widespread acceptance, this examination has been criticized for its lack of reliability and validity and the high cost of its administration.

Reliability is the measure of both the consistency and precision of a testing tool. The three main sources of variability (decreased reliability) in the oral examination process are: 1) examiner related variability; 2) examination related variability; and 3) candidate related variability. In their paper entitled "Poor inter-rater reliability on mock anesthesia oral examinations" in this edition of the Journal, Jacobsohn, Klock, Avidan and the Oral Examination Group present a study which demonstrates poor inter-rater reliability in a mock oral examination context with raters grading in true isolation.[2] Twenty-five residents were examined in a mock examination process resembling the American Board of Anesthesiology (ABA) format on two occasions six weeks apart (E1 and E2). The examinations were videotaped and scored by three experienced ABA examiners and three experienced Royal College of Physicians and Surgeons of Canada (RCPSC) examiners in isolation. The examiners were provided with a standardized scoring system and an educational package to aid with standard setting. The inter-rater reliability as determined by using intraclass correlation coefficients was poor: 0.243 (0.177–0.305) for E1 and 0.405 (0.331–0.470) for E2. For 48% of the candidates examined, the chance of passing or failing was examiner dependent.

Previous studies have demonstrated significantly better inter-rater reliability in the anesthesia oral examination process. Schubert reported inter-rater reliability as generalized reliability coefficients for both final grade received and pass-fail determination on 441 practice oral examinations given to 190 residents using the ABA format.[3] Inter-rater reliability was 0.72 for the final grade received and 0.68 for the pass-fail determination. This compares favourably with the results found by Kearney in a study using a structured oral examination format for practice examinations similar to that currently used by the RCPSC.[4] Twenty faculty examined 26 residents from two Canadian residency programs (sites A and B). Standardized questions were scored using global rating scales with anchored performance criteria. Each candidate was scored by a pair of examiners at the initial session and again subsequently from a videotaped recording. Inter-rater agreement was 0.51 for time 1 and .79 at time 2 for site A, 0.71 at time 1 and .48 at time 2 for site B. These results were classified as fair to good inter-rater reliability and the wide range of correlations found was felt to be due to several study limitations. The residents examined were at different levels of training with 25% presenting for their first practice oral examination. Evidence suggests that examiners are less consistent when rating poor performances.[5] The number of scores per examiner on which correlations were determined was also quite small. Increasing the number of questions per candidate might improve the overall inter-rater reliability.

Some of the same limitations can be found in the current study by Jacobsohn. The residents were from CA1 to CA3 years of residency and the pass rate at both time E1 and time E2 was low. As mentioned pre-

From the Department of Anesthesia,* St. Michael's Hospital, University of Toronto, Toronto, Ontario; and the Department of Anesthesiology and Pain Medicine,† University of Alberta, Edmonton, Alberta, Canada.

Address correspondence to: Dr. Patricia Houston, Department of Anesthesia, St. Michael's Hospital, 30 Bond St., Toronto, Ontario M5B 1W8, Canada. Phone: 416-864-5071; Fax 416-864-6014; E-mail: houstonp@smh.toronto.on.ca

viously, poor performance decreases inter-rater reliability. In addition, a limited number of examiners was used and although a package of educational material was provided to the examiners to aid in standard setting, no effort was made to achieve consensus among the examiners as to what constituted an appropriate response and what level of competency the examination was meant to determine. The RCPSC examiners had no prior knowledge of, or experience in using the ABA scoring system, thus, the lack of proper examiners' training may have contributed to poor reliability. Finally, analysis of the data may have been more appropriately done using a generalizability theory approach which would have allowed the authors to analyze the variance in the scores due to sources other than inter-rater difference such as the exam content and candidate variability.[6]

The second major finding in the study by Jacobsohn was that a teaching intervention targeted at oral examination skills did not significantly improve examination performance. As the inter-rater reliability found using their examination format was low, it is likely that their measurement tool was not capable of detecting such an improvement. However, this study did demonstrate that the exercise of taking practice oral examinations significantly improves examination performance. This was previously reported by Schubert who found that oral practice examination pass rate increased with both anesthesia training duration and greater exposure to the oral practice examination.[3]

No conclusions regarding the reliability or validity of the RCPSC anesthesia certification examinations should be drawn based on the results of Jacobsohn's study. The RCPSC Anesthesia Examination Board has introduced many processes to address these issues. Examiners receive standardized training, candidates at similar levels of competency are examined, standard setting activities are performed to determine the level of competency acceptable for success in the examination process, and each candidate on any one day is examined using the same examination tool. The questions are truly standardized in presentation, a rating scale with performance anchors is used to score performance and model answers are created with which to compare the candidates' answers. The premise on which the conclusions of the present study is based - that examiners read non-verbal signals from each other which influence their scoring and which the authors surmise may have led to higher reliability coefficients observed in the other studies, is based on tenuous information. The study quoted to support this theory is 40 years old. The examination format used at that time bears little resemblance to that currently in place. The RCPSC has an extensive quality assurance process for its examinations. There is ongoing psychometric evaluation of examination results and monitoring of the examination process to ensure standardization of presentation and independence in scoring. The RCPSC Anesthesia Examination Board responds to both changes in knowledge of anesthesiology and changes in educational theory to optimize the reliability and validity of the examination process.

The oral examination remains a popular method of assessment of postgraduate medical trainees. Ideally, the competency of residents could be judged with precision by reliable measures of treatment outcomes or by direct observation of their clinical performance in an objective format acceptable to the candidates, their peers and to the public. Such objective criteria would provide irrefutable evidence of the validity of the examination process. However, neither reliable treatment-outcome indices nor observation formats have been fully developed and validated. Therefore, competence may need to be evaluated using multiple assessment modalities such as written and oral tests complemented by performance-based criteria. Simulation may provide a future method of reliable and valid performance-based evaluation but is not yet well studied or feasible in many centres.[7]

The physicians of the future need to be equipped with the tools and concepts that foster mindful, attentive and effective deliberation and reflection. Further research and refinement of the oral examination and other tools of evaluation may lead to improved assessment of the physician's mastery of these skills. As Eagle stated "It may be that the greatest strength of the oral examination is not as a measurement instrument but as a teaching device".[1] Preparation for the oral examination may lead the candidates not only to a greater command of the knowledge required, but also improve their ability to integrate and communicate ideas and to demonstrate reflective practice.

Le processus d'examen oral - un vrai ou un faux étalon-or ?

Le processus d'examen oral a été une composante traditionnelle de la reconnaissance professionnelle en anesthésie depuis plus de 70 ans.[1] On croit que

l'examen oral peut servir à évaluer l'habileté à résoudre des problèmes, les aptitudes de communication et de collaboration et la compétence à approfondir les connaissances d'un sujet. Malgré son acceptation largement répandue, cet examen a été critiqué pour son manque de fiabilité et de validité et le coût élevé de son administration.

La fiabilité est la mesure de la constance et de la précision d'un outil de test. Les trois principales sources de variabilité (diminution de la fiabilité) du processus d'examen oral sont : 1) la variabilité reliée à l'examinateur, 2) reliée à l'examen et 3) reliée au candidat. Dans leur article intitulé «Le pauvre coefficient d'objectivité d'examens oraux en anesthésie simulés», Jacobsohn, Klock, Avidan et l'*Oral Examination Group* présentent une étude qui démontre un pauvre coefficient d'objectivité dans le contexte d'une simulation d'examen oral avec des examinateurs en isolement complet.[2] Vingt-cinq résidents ont été interrogés à deux reprises et à six semaines d'intervalle (E1 et E2) lors d'un examen simulé ressemblant à celui de l'*American Board of Anesthesiology* (ABA). Les examens ont été enregistrés sur bandes vidéo et notés isolément par trois examinateurs expérimentés de l'ABA et du Collège royal des médecins et chirurgiens du Canada (CRMCC). Les examinateurs possédaient un système de notation standardisé et une trousse pédagogique pour aider à la normalisation. Le coefficient d'objectivité, déterminé par les coefficients de corrélation intraclasses, a été pauvre : 0,243 (0,177–0,305) pour E1 et 0,405 (0,331–0,470) pour E2. Chez 48 % des candidats examinés, la probabilité de succès ou d'échec était liée à l'examinateur.

Des études antérieures ont montré un coefficient d'objectivité significativement meilleur dans le processus d'examen oral en anesthésie. Schubert a décrit le coefficient d'objectivité comme les coefficients d'objectivité généralisés de la note finale reçue et de la détermination du succès ou de l'échec de 441 examens oraux de pratique présentés à 190 résidents dans le format de l'ABA.[3] Le coefficient d'objectivité était de 0,72 pour la note finale reçue et de 0,68 pour la détermination du succès ou de l'échec. Ces résultats se comparent favorablement à ceux de Kearney dans une étude utilisant un examen oral structuré pour des examens de pratique similaires à ceux qui sont couramment utilisés par le CRMCC.[4] Vingt examinateurs ont testé 26 résidents de deux programmes de résidence canadiens (sites A et B). Les questions standardisées ont été notées au moyen d'échelles d'évaluation globale comportant des critères de performance reconnus. Chaque candidat était noté par deux examinateurs à la première session et encore par la suite à partir de l'enregistrement vidéo. La concordance inter-examinateur a été de 0,51 pour le temps 1 et 0,79 au temps 2 pour le site A, 0,71 au temps 1 et 0,48 au temps 2 pour le site B. Ces résultats ont été classés comme étant des coefficients d'objectivité de passables à bons et le grand éventail de corrélations trouvé a été vu comme la conséquence des limites de l'étude. Les résidents testés avaient une formation de niveau différent et 25 % d'entre eux se présentaient pour la première fois à un examen oral simulé. Il est prouvé que les examinateurs sont moins constants à juger des performances pauvres.[5] Le nombre de scores par examinateur sur lesquels les corrélations ont été déterminées était aussi plutôt bas. L'augmentation du nombre de questions par candidat aurait pu améliorer le coefficient général d'objectivité.

Des limites semblables apparaissent dans la présente étude de Jacobsohn. Les résidents étaient de CA1 à CA3 ans de résidence et le taux de réussite aux deux examens, E1 et E2, a été faible. Comme on l'a dit antérieurement, une pauvre performance fait baisser le coefficient d'objectivité. De plus, le nombre d'examinateurs était limité et, même si une trousse de matériel pédagogique leur a été fournie pour faciliter la normalisation, aucun effort n'a été fait pour atteindre un consensus parmi les examinateurs sur ce que constitue une réponse appropriée et sur le niveau de compétence que l'examen était censé déterminer. Les examinateurs du CRMCC n'avaient pas de connaissance antérieure ou d'expérience du système de notation de l'ABA. Le manque de formation appropriée des examinateurs peut donc avoir contribué au peu de fiabilité. Finalement, l'analyse des données aurait été mieux faite en utilisant la théorie de la généralisabilité qui aurait permis d'analyser la variance des scores d'autres sources que la différence inter-examinateur, comme le contenu de l'examen et la variabilité des candidats.[6]

Le second résultat important de l'étude de Jacobsohn était qu'une intervention de formation ciblée sur les techniques de l'examen oral n'a pas amélioré la performance à l'examen de façon significative. Comme le coefficient d'objectivité trouvé en utilisant leur format d'examen a été bas, il est probable que leur outil de mesure ne pouvait détecter cette amélioration. Cependant, l'étude a démontré que l'exercice d'examens oraux simulés améliore significativement la performance à l'examen. Schubert avait antérieurement trouvé que le taux de réussite de l'examen oral simulé a augmenté avec la durée de la formation en anesthésie et la plus grande exposition à l'examen oral simulé.[3]

Aucune conclusion sur la fiabilité ou la validité des examens de reconnaissance professionnelle en anesthésie

du CRMCC ne peut être tracée sur la base des résultats de l'étude de Jacobsohn. Le comité d'examen en anesthésie du CRMCC a présenté de nombreux procédés pour régler ces questions. Les examinateurs reçoivent une formation standardisée, des candidats de niveaux de compétence similaires sont examinés, des activités de normalisation sont réalisées pour déterminer le niveau de compétence acceptable pour le succès du processus d'examen et, pour un jour donné, le même matériel est présenté à chaque candidat. Les questions sont standardisées dans leur présentation, une échelle de notation avec des marqueurs de performance est utilisée pour coter la performance et des réponses types sont créées avec lesquelles on peut comparer les réponses des candidats. L'hypothèse sur laquelle les conclusions de la présente étude sont fondées - que les examinateurs ont perçu des signaux non-verbaux les uns des autres, ce qui a influencé leur notation, et ce que les auteurs ont présumé aurait pu conduire aux coefficients de fiabilité plus élevés observés dans d'autres études - repose sur des renseignements fragiles. L'étude citée pour appuyer cette théorie a 40 ans. Le format de l'examen utilisé à cette époque ressemble peu à celui qui est présentement en vigueur. Le CRMCC a un processus élaboré d'assurance de la qualité pour ses examens. Il y a une évaluation psychométrique permanente des résultats d'examen et une surveillance du processus d'examen pour assurer la standardisation de la présentation et l'indépendance dans la notation. Le comité d'examen en anesthésie du CRMCC répond aux changements dans les connaissances de l'anesthésiologie et dans la théorie pédagogique pour optimaliser la fiabilité et la validité du processus d'examen.

L'examen oral demeure une méthode populaire d'évaluation des études médicales postdoctorales. Idéalement, la compétence des résidents pourrait être jugée avec précision par des mesures fiables des résultats des traitements ou par l'observation directe de la performance clinique dans un format objectif acceptable aux candidats, à leurs pairs et au public. Ces critères objectifs fourniraient une preuve irréfutable de la validité du processus d'examen. Toutefois, ni les indices fiables de résultats de traitement, ni les formats d'observation n'ont été complètement élaborés et validés. Par conséquent, il faudrait peut-être évaluer la compétence selon des modalités multiples comme des tests écrits et oraux complétés par des critères fondés sur la performance. La simulation peut devenir une méthode d'évaluation fiable et valide fondée sur la performance, mais elle n'est pas encore bien étudiée ou applicable dans de nombreux centres.[7]

Les médecins de l'avenir doivent avoir les outils et les concepts qui stimulent la discussion et la réflexion conscientes, attentives et efficaces. D'autres recherches et le raffinement de l'examen oral ainsi que d'autres outils pourront améliorer l'évaluation de la maîtrise qu'ont les médecins de ces techniques. Comme l'a déclaré Eagle «La plus grande force de l'examen oral n'est peut-être pas en tant qu'instrument de mesure, mais en tant que technique d'enseignement».[1] La préparation à l'examen oral peut mener les candidats non seulement à une plus grande maîtrise du savoir requis, mais aussi améliorer leur habileté à intégrer et à communiquer des idées et à démontrer une pratique réfléchie.

References

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

2 *Jacobsohn E, Klock PA, Avidan M; Oral Examinations Group.* Poor inter-rater reliability on mock anesthesia oral examinations. Can J Anesth 2006; 53: 659–68.

3 *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: 288–98.

4 *Kearney RA, Puchalski SA, Yang HY, Skakun EN.* The inter-rater and intra-rater reliability of a new Canadian oral examination format in anesthesia is fair to good. Can J Anesth 2002; 49: 232–6.

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

6 *Brennan RL, Johnson EG.* Generalizability of performance assessments. Educational Measurement: Issues and Practice 1995; 14: 9–12.

7 *Savoldelli GL, Naik VN, Joo HS, Houston PL, Graham M, Yee B, Hamstra SJ.* Evaluation of patient simulator performance as an adjunct to the oral examination for senior anesthesia residents. Anesthesiology 2006; 104: 475–81.