Simulation performance checklist generation using the Delphi technique

[Génération d'une liste de vérification de la performance simulée à l'aide de la méthode Delphi]

Pamela J. Morgan MD CCFP FRCPC,*† Jenny Lam-McCulloch MSc,† Jodi Herold-McIlroy PhD,‡ Jordan Tarshis MD FRCPC†

Purpose: Performance assessment using high fidelity simulation is problematic, due to the difficulty in developing valid and reliable evaluation tools. The Delphi technique is a consensus based content generation method used for multiple purposes such as policy development, best-evidence practice guidelines and competency assessments. The purpose of this study was to develop checklists using a modified Delphi technique to evaluate the performance of practicing anesthesiologists managing two simulated scenarios.

Methods: The templates for two simulation scenarios were emailed to five anesthesiologists who were asked to generate performance items. Data were collated anonymously and returned. An *a priori* decision was made to delete items endorsed by ≤ 20% of participants. This process of collection, collation and re-evaluation was repeated until consensus was reached. Four independent raters used the checklist to assess three subjects managing the two simulation scenarios. Inter-rater reliability was assessed using average measures intraclass correlation (ICC) and repeated measures analysis of variance (ANOVA) was used to assess differences in difficulty between scenarios.

Results: The final checklists included 131 items for scenario 1 and 126 items for scenario 2. The mean inter-rater reliability was 0.921 for scenario 1 and 0.903 for scenario 2. Repeated measures ANOVA revealed no statistically significant difference in difficulty between scenarios.

Discussion: The Delphi technique can be very useful to generate consensus based evaluation tools with high content and face validity compared to subjective evaluative tools. Since there was no difference in scenario difficulty, these scenarios can be used to determine the effect of educational interventions on performance.

CAN J ANESTH 2007 / 54: 12 / pp 992–997

Objectif *: L'évaluation de la performance est problématique lorsqu'on a recours à une simulation de haute fidélité, ceci en raison de difficultés rencontrées lors du développement d'outils d'évaluation valables et fiables. La méthode Delphi est une méthode de génération de contenu qui se base sur un consensus ; elle est utilisée dans divers contextes tels que le développement de directives, des guides de pratique basés sur les meilleures données probantes, et l'évaluation des compétences. L'objectif de cette étude était de développer des listes de vérification en utilisant une méthode Delphi modifiée et ce, afin d'évaluer la performance des anesthésiologistes actifs gérant deux scénarios simulés.*

Méthode *: Les modèles pour deux scénarios de simulation ont été envoyés par courriel à cinq anesthésiologistes, auxquels on a demandé de générer des rubriques de performance. Les données ont été rassemblées de façon anonyme et renvoyées. Une décision a priori a été prise d'effacer les rubriques approuvées par ≤ 20 % des participants. Ce processus de récolte, de comparaison et de réévaluation a été répété jusqu'à ce que l'on atteigne un consensus. Quatre évaluateurs indépendants ont utilisé la liste de contrôle pour évaluer trois sujets prenant en charge les deux scénarios de simulation. La crédibilité inter-évaluateurs a été évaluée à l'aide de mesures de corrélation intraclasse (ICC) moyennes et des mesures répétées de l'analyse de variance (ANOVA) ont été utilisées afin d'évaluer les différences de difficulté entre les scénarios.*

Résultats *: Les listes de contrôle finales comprenaient 131 rubriques pour le scénario no. 1 et 126 rubriques pour le scénario no. 2. La crédibilité inter-évaluateurs moyenne était de 0,921 pour le scénario no. 1 et de 0,903 pour le scénario no. 2. Les mesures répétées ANOVA n'ont pas révélé de différence statistiquement significative de la difficulté entre les scénarios.*

Discussion *: La méthode Delphi peut être très utile pour générer des outils d'évaluation basés sur un consensus avec un contenu élevé et une validité apparente par rapport à des outils d'évaluation*

From the Departments of Anesthesia, Women's College Hospital;* Sunnybrook Health Sciences Centre;† and the Wilson Centre for Research in Education,‡ University of Toronto, Toronto, Ontario, Canada.

Address correspondence to: Dr. P. J. Morgan Department of Anesthesia, Women`s College Hospital, 76 Grenville St. Toronto, Ontario M5S 1B2, Canada. Phone: 416-323-6400, ext. 4087; Fax: 416-323-6307; E-mail: pam.morgan@utoronto.ca

This study was supported by a grant from the Canadian Institutes of Health Research. The authors have no commercial or non-commercial affiliations, conflict of interest with this work nor any other associations such as consultancies.

Accepted for publication April 26, 2007.
Revision accepted July 9, 2007.
Final revision accepted September 3, 2007.

subjectifs. Etant donné qu'il n'y a pas eu de différence de difficulté entre les scénarios, ces derniers peuvent être utilisés pour déterminer l'effet d'interventions éducationnelles sur la performance.

PERFORMANCE assessment using high fidelity simulation is problematic, largely due to the difficulty in developing valid and reliable evaluation tools. Checklists are easy to use and reliable but appropriate content is essential for validity. The Delphi technique is a consensus based content generation method used for multiple purposes such as policy development, best-evidence practice guidelines and competency assessments.[1] It involves item generation/endorsement, iteration and feedback, and does not require face-to-face meetings.

A number of features separate the Delphi technique from other data collection methodologies. Firstly, the collection of data is anonymous which attempts to reduce the influence of factors such as group conformity, power, and influences of others on responses. The objective of anonymity is to allow introduction and evaluation of ideas and concepts by removing some of the common biases normally occurring in the face-to-face group. If an idea is unsuitable, no one loses face from having been the individual to introduce it. Votes are more frequently changed when the identity of a given voter is not available.

The modified Delphi approach consists of providing an initial parameter with or without some context in which to reply[2] and this process has been used widely to develop policies,[3] identify priorities[4] and the necessity of medical procedures.[5] The ease of applying this technique using a computer-based program has shifted focus from the traditional pen and pencil to electronically generated Delphi responses.

The purpose of this study was to develop a performance checklist for practicing anesthesiologists managing two simulated patient scenarios using a modified computer-based Delphi technique. The secondary objectives were to pilot test the newly developed checklist with four raters and determine inter-rater reliability and ease of use. Finally, this study sought to determine any differences in difficulty between the two scenarios.

Methods

After Research Ethics Board approval, the outline of two anesthetic scenarios were developed by research team consensus and included the management of a patient undergoing anesthesia for a laparoscopic cholecystectomy and a second scenario of a patient undergoing an urgent laparotomy for large bowel obstruction. Within the context of each scenario, there were two critical events, one of which required the application of an advanced cardiac life support (ACLS) algorithm. As well, situations were introduced into the scenario that would enhance the likelihood of an anesthesiologist's error. Examples included missing pieces of information from the patient chart, or missing anesthetic equipment or drugs. The scenario was broken down into nine segments which were labelled: 1) Preoperative Assessment; 2) Machine/Equipment; 3) Drugs/Medications; 4) Monitors; 5) Induction of Anesthesia; 6) Maintenance Phase I; 7) Critical Event 1; 8) Critical Event 2; and 9) Management of an ACLS Event. This information was entered into a Microsoft Excel spreadsheet leaving lines for reviewers to complete after each segment title. A column asking for the weighting of the item was added on the far right of the spreadsheet.

In the first round, five anesthesiologists from two independent academic centres were recruited for participation in the study. Each participant was emailed the history, physical findings, laboratory findings and "missing" information for each patient scenario. As well, they were emailed the scenario template on an excel spreadsheet. Each participant was asked to insert the specific items that an anaesthesiologist should be expected to perform on encountering the scenario event. In addition, they were asked to complete an error weighting for each item and enter it into the appropriate column. Item error weighting was defined as follows: Weighting: 1) No risk: an incident without any potential risk for the patient; 2) Low risk: an incident which could have led to reversible damage to the patient; 3) Medium risk: an incident which could have led to irreversible damage to the patient; and 4) High risk: a potentially fatal incident.[6]

Once the checklists were returned, they were anonymously collated and emailed back to the participants in round 2. Participants were then asked to check off whether the item should remain or be deleted, and to assign a weighting if it was a new item to them or re-assign a weighting if deemed appropriate. This process was repeated until no further items were added or deleted and no further changes were made in weighting of items. An *a priori* decision was made to delete all responses endorsed by less than ≤ 20% of respondents. For items whose weightings differed, a median weighting was determined. For the section where ACLS guidelines should be followed, the investigators compared the responses and ensured that they conformed with current ACLS guidelines.

TABLE I Percentage weightings of scenario items

Weighting	*Scenario 1 (n = 104)**	*Scenario 2 (n = 99)**
1: No risk: an incident without any potential risk for the patient	10 (9.62%)	12 (12.12%)
2: Low risk: an incident which could have led to reversible damage to the patient	38 (36.54%)	35 (35.35%)
3: Medium risk: an incident which could have led to irreversible damage to the patient	50 (48.08%)	47 (47.47%)
4: High risk: a potentially fatal incident	6 (5.77%)	5 (5.05%)

*Not including unplanned errors.

Previous experience with simulation research identified "unplanned" errors that occurred during simulation scenarios. Therefore, for each of the nine sections of the checklist, three lines were included so that raters could add, identify and weight unplanned errors committed by the subjects. These additions added 27 items to the final total.

Once the checklists were finalized, three anesthesiologists were recruited to manage the simulated scenarios at two independent simulation centres. Their performances were videotaped and the videotapes sent to four academic anesthesiologists not involved in the first phase of the study. The anesthesiologists were asked to rate the simulated performances using the newly developed checklists and to provide written feedback regarding the ease of use of the checklists. Comparison of scores of the two scenarios was done to determine if there was a significant difference in scenario difficulty.

Statistics

Descriptive statistics were used to analyze the checklist development, specifically the proportion of reviewers endorsing each item. Inter-rater reliability of the four raters rating the performance of one given subject in a given scenario was analyzed using the average measures intraclass correlation coefficient (ICC) for dichotomous data. A repeated measures analysis of variance (ANOVA) with scenario and rater as the within-subjects factor was used to determine whether there was any difference in average scores (i.e., difficulty) between scenarios. Analysis was done using SPSS Version 13.0 (SPSS Inc., Chicago, IL, USA).

Results

All five anesthesiologists who were recruited to participate, completed the development of two performance checklists using a computer-based modified Delphi technique. On the first iteration, 78 items were generated for scenario 1 and 83 items for scenario 2. In round 2, two items were deleted due to an endorsement of ≤ 20%. In total after round 2, 31 items were added when both scenarios were considered together. In round 3, nine further items were added and one was deleted. In the final round, two further items were added after which no further items were added or deleted, nor were there any further alterations in weighting. After expert review and correlation with current ACLS guidelines, 12 items were added to the management of the ACLS algorithms. The final checklist then consisted of 104 items for scenario 1 and 99 items for scenario 2. With the addition of the 27 unplanned error items, the totals for each scenario checklist were 131 for scenario 1 and 126 for scenario 2 (Appendix, available as Additional Material at www.cja-jca.org). The percentage weightings of the checklist items (not including the unplanned error items) are found in Table I.

Four independent raters completed the performance checklists for each of three subjects managing the two scenarios. There were varying degrees to which all four raters completed all checklist items, primarily due to difficulties seeing some components on the videotapes. The average number of items completed by all four raters for scenario 1 was 76 and for scenario 2 was 70; the majority of the remaining items were completed by three of the four in most circumstances. There was good inter-rater reliability on both individual scenarios

TABLE II Inter-rater reliability of four raters using checklists to assess videotaped performance

	Scenario 1			*Scenario 2*		
	*ICC**	*95% CI***	*# Items****	*ICC**	*95%CI*	*# Items****
Subject 1	.915	(.881, .942)	78	.878	(.827, .917)	80
Subject 2	.924	(.894, .948)	89	.892	(.846, .926)	77
Subject 3	.923	(.881, .948)	66	.939	(.905, .963)	51
Average ICC	.921	(.911, .930)		.903	(.840, .966)	

*Average measures intraclass correlation coefficient (ICC) based on four raters; **95% confidence interval (CI) for estimate of ICC; ***Number of items with no missing data across four raters. Scenarios 1 and 2 had 104 and 99 items respectively.

as demonstrated by the high average measure intraclass correlation coefficients (Table II). These coefficients were estimated on the basis of the items for which data were available from all four raters.

Total scores were computed for each subject's performance in each of the two scenarios using both weightings for the items as generated by the experts, and without item weightings. The average score for the three subjects as rated by the four raters was 53.42 (SD = 6.86) on scenario 1 and 49.50 (SD = 8.57) on scenario 2. The mean difference score between scenarios for each subject was 3.91 (SD = 3.39). Repeated measures ANOVA with scenario and rater as within-subjects factors using the unweighted scores as the dependent measure demonstrated no difference in difficulty of scenarios.

Comments from video reviewers indicated that overall, the checklists were easy to use. Some items were hard to assess using the videotape because of the position of the camera and the anesthesiologist. Many of these items involved the observation of the anesthesiologist's machine check. The most common machine items which could not be assessed using the videotapes were, in descending order: a) scavenging system working, b) oxygen analyzer functioning, c) unidirectional valve and soda lime check, and d) vaporizer filled, ports closed. In addition, the item which requested "appropriate response to distraction at time of critical event" was also felt to be difficult to assess.

Discussion

The Delphi technique was developed by the RAND Corporation in the 1950s providing a quantitative means of assessing the judgment of experts.[A] Multiple rounds of anonymous responses are collated and returned to the users who send new responses depending on the changes. This process is iterated until the opinions have converged or no further additions/deletions are evident. There is however, no universally agreed upon guideline for its use, nor standardization of methodology.[7] In addition, the number of experts to be recruited appears in many cases, to be based on "common sense and practical logistics" and may vary between five and 154.[7,8]

In this study, five academic anesthesiologists were invited to participate in the generation of performance checklists for practicing anesthesiologists. Due to the nature of the study, a purposive sampling was undertaken since the individuals required expertise in the specialty of anesthesia to complete the performance items.[9] With purposive sampling, individuals are not randomly selected but are "hand picked" by the research team. Although the number of actual "experts" required to be involved in the Delphi technique varies widely, it is important to address the endpoint in question when considering the number of experts to recruit.[1,9] When qualitative data is required for generation of policies and guidelines to practice, a larger number of subjects might be warranted. However, a number of issues arise as numbers of experts increase. Firstly, it has been well demonstrated that response rate drops as expert numbers increase and secondly, data management can become unwieldy.[9,10] In this study, we needed anesthesiologists involved in education who could produce expected performance items for two simulated scenarios. Due to the nature of the performance items, it was not anticipated that wide variations in opinions would exist, considering that much of what would be expected already exists as published guidelines i.e., anesthesia machine check[B] and ACLS guidelines. In addition, due to the time and financial requirements to complete an unknown number of iterations, we decided that to optimize return of checklists, a smaller rather than larger number of experts would be recruited.

There are differing Delphi formats with the "traditional" method using the literature or focus groups to generate content for the first round.[11] In this study, it was necessary to provide the experts with a skeleton of the scenario to be managed in order for them to develop items of importance, thereby classifying the methodology as a modified Delphi technique.

Consensus was considered to have been achieved when no further performance items were listed, nor were any suggestions for deletion or alterations made. After the final round of data collection and collation, we deleted performance items endorsed by ≤ 20%. This number has been suggested as an appropriate percentage when developing health measurement scales[12] and for Delphi responses.[13] We required only three iterations following the initial template development to achieve 'consensus' which is consistent with other studies in the literature.[13,14] The fact that only three iterations were required also supports the "stability" of the responses which some have suggested to

A *Fitch K, Bernstein S, Aguilar M, et al.* The Rand/UCLA Appropriateness Method User's Manual. Santa Monica, CA: Rand Corporation; 2001.

B http://www.cas.ca/members/sign_in/guidelines/practice_of_anesthesia/default.asp?load=appendix_iii (accessed September 2007).

be a reliable indicator of consensus.[9]

When considering the importance of high-stakes performance based examinations, it is critical that performance tools have adequate validity and reliability.[15–17] The Delphi technique assures that the identified items have high face validity and when consensus is achieved, concurrent validity is evident.[18,19] Inter-rater reliability using the newly developed checklists was high, mirroring the results of other simulation studies.[17,20] Future studies with more subjects will be conducted using a generalizability framework for analysis.

When using the Delphi technique, it must be kept in mind that one is achieving consensus but not necessarily the "correct" answer.[21] This technique forces participants to reach a consensus and does not allow conversation to discuss issues. Therefore, one relies on the 'expertise' of the participants. In this study, there were no apparent disagreements between the experts as to the inclusion or exclusion of any item.

It is important, before implementation, to pilot test any newly developed tool in order to determine feasibility and ease of use.[22] Ultimately, these newly developed performance checklists will be used to evaluate the performance of practicing anesthesiologists in simulated anesthetic scenarios. In this study, it was found that the same three subjects' scores on both scenarios were not statistically different reflecting a similar complexity or "difficulty" of content. In the future, therefore, these two scenarios could be used interchangeably to compare the effect of educational interventions on performance.

Videotapes of performance will be used by raters to complete the checklists. The final performance checklists consisted of 131 (scenario 1) and 126 (scenario 2) items respectively. Although the reviewers who piloted the use of the checklists in this study did not find the number of items too cumbersome, some aspects of performance were difficult to view due to camera position and therefore some items were impossible to score using the checklists. More sophisticated filming equipment that allows for multiple views of the operating room may facilitate identification of all items. Alternatively, it may be necessary to have "live" raters completing the checklists for certain sections of the simulated scenarios that have been identified as impossible to identify on viewing the videotaped performance or if items are still unable to be viewed, they may ultimately have to be deleted.

A limitation of our study was the fact that we did not compare our final checklist using the Delphi technique with any other method of developing the tool. It may have been possible to develop a tool by direct consultation with experts during a single meeting. While this method is easier and perhaps equally reliable, we felt that the effort of using the Delphi technique in generating a final performance checklist was warranted due to the proposed use of the checklists in the future assessment of anesthesiologists' performances in the simulated environment.

In conclusion, the Delphi technique offers the means to acquire consensus for performance-based assessments in an anonymous way that maximizes the face, content and concurrent validity of the tool. These performance assessment tools can now be used to evaluate performance during high-fidelity simulation scenarios and to compare the effects of educational interventions on simulation performance which may ultimately improve patient safety.

References

1 *Clayton M.* Delphi: a technique to harness expert opinion for critical decision-making tasks in education. Educ Psychol 1997; 17: 373–86.

2 *Martino J.* Technological Forecasting for Decision Making, 2nd ed. New York: North-Holland; 1983.

3. *Lavis JN, Anderson GM.* Appropriateness in health care delivery: definitions, measurement and policy implications. CMAJ 1996; 154: 321–8.

4 *Brook RH, Kamberg CJ.* Appropriateness of the use of cardiovascular procedures: a method and results of this application. Schweiz Med Wochenschr 1993; 123: 249–53.

5 *Kahan J, Bernstein SJ, Leape LL, et al.* Measuring the necessity of medical procedures. Med Care 1994; 32: 357–65.

6 *Chopra V, Bovill JG, Spierdijk J, Koornneef F.* Reported significant observations during anaesthesia: a prospective analysis over an 18-month period. Br J Anaesth 1992; 68: 13–7.

7 *Evans C.* The use of consensus methods and expert panels in pharmacoeconomic studies. Practical applications and methodological shortcomings. Pharmacoeconomics 1997; 12: 121–9.

8 *Keeney S, Hasson F, McKenna H.* Consulting the oracle: ten lessons from using the Delphi technique in nursing research. J Adv Nurs 2006; 53: 205–12.

9 *Hassan F, Keeney S, McKenna H.* Research guidelines for the Delphi survey technique. J Adv Nurs 2000; 32: 1008–15.

10 *McKenna HP.* The essential elements of a practitioners' nursing model: a survey of clinical psychiatric nurse managers. J Adv Nurs 1994; 19: 870–7.

11 *Kearney RA.* Defining professionalism in anaesthesiology. Med Educ 2005; 39: 769–76.

12 *Streiner D.* Selecting the items. *In*: Streiner DL, Norman GR (Eds). Health Measurement Scales. A

Practical Guide to Their Development and Use, 2nd ed. Oxford: Oxford University Press; 1995: 54–68.
13 *Green B, Jones M, Hughes D, Williams A*. Applying the Delphi technique in a study of GPs' information requirements. Health Soc Care Community 1999; 7: 198–205.
14 *Beech B*. Studying the future: a Delphi survey of how multi-disciplinary clinical staff view the likely development of two community mental health centres over the course of the next two years. J Adv Nurs 1997; 25: 331–8.
15 *Devitt JH, Kurrek MM, Cohen MM, Cleave-Hogg D*. The validity of performance assessments using simulation. Anesthesiology 2001; 95: 36–42.
16 *Morgan PJ, Cleave-Hogg D, DeSousa S, Tarshis J*. High-fidelity patient simulation: validation of performance checklists. Br J Anaesth 2004; 92: 388–92.
17 *Morgan PJ, Cleave-Hogg DM, Guest CB, Herold J*. Validity and reliability of undergraduate performance assessments in an anesthesia simulator. Can J Anesth 2001; 48: 225–33.
18 *Williams PL, Webb C*. The Delphi technique: a methodological discussion. J Adv Nurs 1994; 19: 180–6.
19 *Streiner DL, Norman GR*. Validity. *In*: Streiner DL, Norman GR (Eds). Health Measurement Scales: A Practical Guide to Their Development and Use, 2nd ed. Oxford: Oxford University Press; 1995: 144–62.
20 *Devitt JH, Kurrek MM, Cohen MM, et al*. Testing the raters: inter-rater reliability of standardized anaesthesia simulator performance. Can J Anaesth 1997; 44: 924–8.
21 *Keeney S, Hasson F, McKenna HP*. A critical review of the Delphi technique as a research methodology for nursing. Int J Nurs Stud 2001; 38: 195–200.
22 *Jairath N, Weinstein J*. The Delphi methodology (part one): a useful administrative approach. Can J Nurs Adm 1994; 7: 29–42.