Comparison of an Aggregate Scoring Method With a Consensus Scoring Method in a Measure of Clinical Reasoning Capacity

Bernard Charlin, Martin Desaulniers, and Robert Gagnon
Faculty of Medicine
University of Montreal
Montreal, Quebec, Canada

Daniel Blouin
Department of Gynecology and Obstetrics
University of Sherbrooke
Sherbrooke, Quebec, Canada

Cees van der Vleuten
Department of Educational Development and Research
University of Maastricht
Maastricht, The Netherlands

Background: *Diversity of clinical reasoning paths of thought among experts is well known. Nevertheless, in written clinical reasoning assessment, the common practice is to ask experts to reach a consensus on each item and to assess students on a unique "good answer."*
Purposes: *To explore the effects of taking the variability of experts answers into account in a method of clinical reasoning assessment based on authentic tasks: the Script Concordance Test.*
Methods: *Two different methods were used to build answer keys. The first incorporated variability among a group of experts (criterion experts) through an aggregate scoring method. The second was made with the consensus obtained from the group of criterion experts for each answer. Scores obtained with the two methods by students and another group of experts (tested experts) were compared. The domain of assessment was gynecology–obstetric clinical knowledge. The sample consisted of 150 clerkship students and seven other experts (tested experts).*
Results: *In a context of authentic tasks, experts' answers on items varied substantially. Amazingly, 59% of answers given individually by criterion group experts differed from the answer they provided when they were asked in a group to provide the "good answer" required from students. The aggregate scoring method showed several advantages and was more sensitive to detecting expertise.*
Conclusions: *The findings suggest that, in assessment of complex performance in ill-defined situations, the usual practice of asking experts to reach a consensus on each item reduces and hinders the detection of expertise. If these results are confirmed by other researches, this practice should be reconsidered.*

Teaching and Learning in Medicine, *14*(3), 150-156

In similar clinical situations, physicians do not collect exactly the same data and do not follow the same paths of thought, even if they obtain the same diagnostic outcome.[1] Furthermore, physicians show substantial variation in performance on any particular real or simulated case.[2,3] These research findings have brought educators to abandon the traditional approach of clinical reasoning teaching and use methods that foster the acquisition of the reasoning processes that physicians really use in practice.[4–6]

This research project has been funded by a grant from the Medical Council of Canada. We thank Martine Fortier and Marie Josée Dupuis for their valuable help in participant recruitment and data collection.

Correspondence may be sent to Bernard Charlin, URDESS, Faculty de Medicine-Direction, Universite de Montreal, CP 6128, succursale Centre-ville, Montreal, Quebec, H3C 3J7 Canada. E-mail: charlinb@meddir.umontreal.ca

Unfortunately, assessment practice has not yet incorporated these findings. We often continue to require a group of experts to provide a common answer for each item in the examination when constructing an answer key for a clinical simulation examination. Items then are removed from tests when consensus is not reached. This may prevent the assessment of situations that belong to authentic clinical life. We also continue to demand that students provide "the good answers" although we know that in practice there is rarely such thing as an absolute good answer.

In clinical simulations used for testing purposes, it is agreed that the diversity of opinion among experts precludes key development by a single individual and that the judgments of several experts are required to develop the answer key.[7] Variations among experts about what constitutes optimal patient management are well documented.[8,9] Until now, although there have been some proposals to account for these variations with an aggregate scoring method,[10,11] the common practice remains to ask experts to reach a consensus.

The Script Concordance Test (SCT) is a new tool of clinical reasoning assessment[12] that stems from current theories and empirical findings about clinical reasoning.[13–15] The SCT is an assessment technique that assesses how knowledge is organized to make adequate decisions in the process of clinical reasoning in comparison to organization of knowledge observed among a group of experts. It does not focus on the outcome of the process, but rather it assesses the microdecisions that are made within that process. From examinee scores, inferences are made on the degree of knowledge organization required to successfully address problems in the assessed domain.

The SCT has a rich authentic context and is case based. Items are made from the questions and actions that physicians ask and make in clinical practice. Examinees are required to make diagnostic, investigative, or therapeutic decisions, when specific elements of information are provided (see Figure 1). The SCT is relatively easy to construct and to administer, and a series of empirical studies have documented that it has a good reliability as well as good face and construct validity.[16–18]

If you were thinking of	And then you find	This hypothesis becomes
(a diagnostic hypothesis)	*(a new clinical information, an imaging study or a laboratory test result)*	–3 –2 –1 0 +1 +2 +3

Figure 1. *Format of items used for diagnostic knowledge assessment.* Note: –3 = *ruled out*; –2 = *much less probable*; –1 = *a little less probable*; 0 = *no effect on this hypothesis*; +1 = *a little more probable*; +2 = *much more probable*; +3 = *certain*.

The tool uses an aggregate scoring method.[10,11] It reflects the variability that experts demonstrate when they answer complex questions belonging to clinical reasoning. Scores on each item are derived from the answers given by a criterion group of experts. The underlying principle is that the answer of any expert reflects a valid opinion that should be taken into account, and answers with poor agreement among experts should not be discarded. Each possible answer to an item receives a partial credit that reflects the number of experts who have given that answer. The scoring process is described in detail subsequently.

In this study, we compared the effect on scores obtained by students and experts with the aggregate method (A method) with the common method in which experts are asked to provide a consensus for each item (C method). The study bears on the gynecology–obstetric domain.

Our research questions were as follows: Do experts provide the same answer when they take the test individually and when they provide "the good answer" in group meeting? What is the distribution of scores, and what are reliability coefficients with both methods? Is one method better than the other to detect expertise among a sample of examinees?

Methods

Setting and Participants

In the school in which the study took place, classes consist of approximately 90 students. Data were collected over a period of a year and a half. Hence, the sample contained 150 students. Students perform rotations during their clerkship in groups of 12 or 13. At the end of each rotation in gynecology–obstetrics, they have an objective structured clinical examination (OSCE). Students were asked to voluntarily take the SCT after the OSCE, and all agreed. The seven staff members of the gynecology–obstetrics department of the school were asked to be the criterion group. Seven other experts (called hereafter "tested experts") from another school were asked to participate in the study. All agreed to participate. Each completed the test individually. The performance of the second group of experts is compared with that of the 150 students.

Construction of Scale

The original SCT contained 60 items (see Figure 2) grouped in three different clinical situations (first trimester bleeding, abnormal uterine bleeding, and request for a contraceptive method). Items were constructed according to the methodology described by Charlin et al.[12] An item analysis was done with the consensus scoring method. We used the iterative

If you were thinking of	And then the patient reports Or you find upon Clinical examination	This hypothesis becomes
14 - Hydatiform mole	Fetal heart is heard on doppler	-3 -2 -1 0 +1 +2 +3
15- Abortion	Enlarged ovaries (± 8cm) on each side	-3 -2 -1 0 +1 +2 +3
16- Abortion	The patient has taken antibiotics a week ago for pneumonia.	-3 -2 -1 0 +1 +2 +3

Figure 2. *Example of items from the test.*

		-3	-2	-1	0	+1	+2	+3
Item 14	Experts	2	5					
	A method	0.40	1	0	0	0	0	0
	Students (N)	118	20	4	1	1	1	5
Item 32	Experts			1		4	2	
	A method	0	0	0.25	0	1	0.50	0
	Students (N)		4	9	1	52	79	5
Item 49	Experts			2		3	2	
	A method	0	0	0.66	0	1	0.66	0
	Students (N)	24	37	31	3	25	27	3
Item 54	Experts		1	3		2	1	
	A method	0	0.33	1	0	0.66	0.33	0
	Students (N)	4	33	41	3	43	24	2
Item 60	Experts	1	1	5				
	A method	0.20	0.20	1	0	0	0	0
	Students (N)	16	33	35	58	5	3	0

Figure 3. *Examples of the scoring system, with the distribution of students' answers. Darkened choices represent the consensus experts made one year later.*

method of deleting items with negative or very small correlation and recalculating alpha coefficient until no more gain on reliability was observed. Fifteen items were removed, leading to a final test composed of 45 items.

Construction of Answer Keys

In a first step, experts from the criterion group were asked to complete the test individually. Their individual answers were used to build the answer key of

method A. In this study we modified the original scoring process used in previous research on the SCT. In the original scoring process, items did not have the same maximum value. This was considered disturbing both by students and by assessment specialists. With the new method, all items get a maximum score of 1. For each item, answers are assigned a weight corresponding to the proportion of the experts who selected it. Credits for each answer then are transformed proportionally to get a maximum score of 1 for modal experts' choices on each item, other experts' choices receiving a partial credit. Answers not chosen by experts receive zero. In the transformation, all scores of an item are multiplied by the total number of experts and divided by the modal value for the item. For example (see Figure 3), if on an item four experts (out of seven) choose Response 1, this choice receives 1 point (4 divided by 4). If two experts choose response 2, this choice receives 0.5 point (2 divided by 4), and if one expert chooses response 3, this choice receives 0.25 point (1 divided by 4). The total score for the test is the sum of credit obtained on each item. Numbers then are transformed to get a maximum of 100.

In a second step, 1 year later, members of the criterion group were asked to meet and provide the best answer by consensus for each item. This scoring method is called the consensus method (C method). Figure 3 illustrates the effect of the two scoring methods on five items of the test.

Statistical Analysis

To answer the first research question, we computed the percentage of answers given individually by criterion experts that differ from "the good answer" provided in a consensus meeting.

To answer the second research question (distribution of scores and reliability coefficients), scores obtained by examinees (i.e., students plus tested experts) with the two methods were studied with descriptive statistics. Cronbach's alpha coefficients were computed with the two answer keys. All statistical comparison were two-tailed and considered significant at 5% alpha level. A Pearson correlation coefficient was calculated to quantify the relation between the two methods. A scatter plot was drawn to visually detect the distribution of examinee scores.

To answer the third question (ability of each method to detect expertise among examinees), we compared the means obtained by students and tested experts with Student's *t* tests. Effect sizes were calculated by using the difference between the groups divided by the standard deviation of students groups.

Results

Research Question 1

Fifty-nine percent of answers given by experts from the criterion group when they answered individually differed from consensual answers (Figure 3 illustrates this variation).

Research Question 2

Descriptive statistics of scores of the 157 examinees (students and experts; see Table 1) show that the range was greater for the C method, with a lower minimum value. The maximum and mean scores were higher with the A method. Skewness and kurtosis were in the normal range for both methods. With method C, the method used for optimization of items, Cronbach alpha was .63. Its value was .52 with the A method. The value of the Pearson correlation coefficient for students (0.72, $p < .001$) indicates that the two methods induced differences in the classification of students. The scatter plot of two sets of scores of students is graphed in Figure 4. Difference in classification was observed mainly for students in the center of the distribution. Ranking of students with more extreme scores showed fewer discrepancies.

Table 1. *Descriptive Statistics With the Two Methods*

	Students		Tested Experts		Total	
Statistics	**A Method**	**C Method**	**A Method**	**C Method**	**A Method**	**C Method**
N	150	150	7	7	157	157
M	54.06	38.73	64.14	46.57	54.51	39.08
SD	8.02	10.79	3.29	6.16	8.14	10.74
Range	47.00	65.00	9.00	18.00	47.00	65.00
Minimum	25.00	4.00	60.00	40.00	25.00	4.00
Maximum	72.00	69.00	69.00	58.00	72.00	69.00
Kurtosis	0.84	0.56	–0.95	1.03	0.75	0.60
Skewness	–0.63	–0.36	0.53	1.16	–0.63	–0.40

Note: The maximum theoretical value for the test is 100.

Research Question 3

Figure 5 depicts the differences obtained with the two methods for the groups of students and tested experts. The A method provided higher scores to tested experts and allowed a better discrimination of scores among examinees. The mean difference was statistically significant, $t(155) = 10.1$, $p < .001$ (0.05-tailed), $d = 17.6$. With the C method, the mean difference between groups was not statistically significant, $t(155) = 7.8$, $p = .06$ (0.05-tailed), $d = 15.4$. The effect size was also different between the two methods (A = 1.25, C = 0.72).

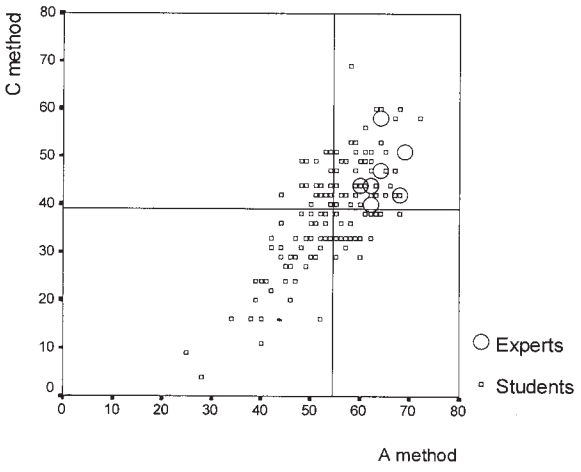

Figure 4. *Scatter plot of A and C scores. Some students who are above the mean with a method are below the mean with other, sometimes wide discrepancies.*

Discussion

In assessment practice, there is a strong tendency toward more authenticity, and the SCT is in line with this tendency. It presents real-life situations to examinees and uses an innovative format to ask them questions that experts ask themselves in the same situations. The goal of this study was to compare two scoring methods. The C method follows the usual practice of asking a group of experts (criterion experts) to provide "the good answer" on each item. The A method considers that criterion experts should pass the test as examinees would do and that each answer given by an expert has an intrinsic value. In constructing the final scale and in the analyses, we optimized items on the C method. Hence, we gave an intrinsic disadvantage to the A method in subsequent analyses, but the reasoning for that approach is that if we found method A to be superior, that finding would have more implications. This explains why the Cronbach alpha coefficient (.63 for the C method) decreased with the A method (.52). If optimization had been done with the A method, the reverse might have been found.

We already knew that in similar clinical situations, physicians do not collect exactly the same data and do not follow the same paths of thought.[1] Thus, the observation that experts' answers vary depending on the context—individual response versus consensus response—is not a surprise. Another finding has potentially strong implications. Fifty-nine percent of answers given by experts from the criterion group were different that those they gave when they placed in the consensus condition. Item number 49 in Figure 3 is an example of this behavior change. Experts' individual

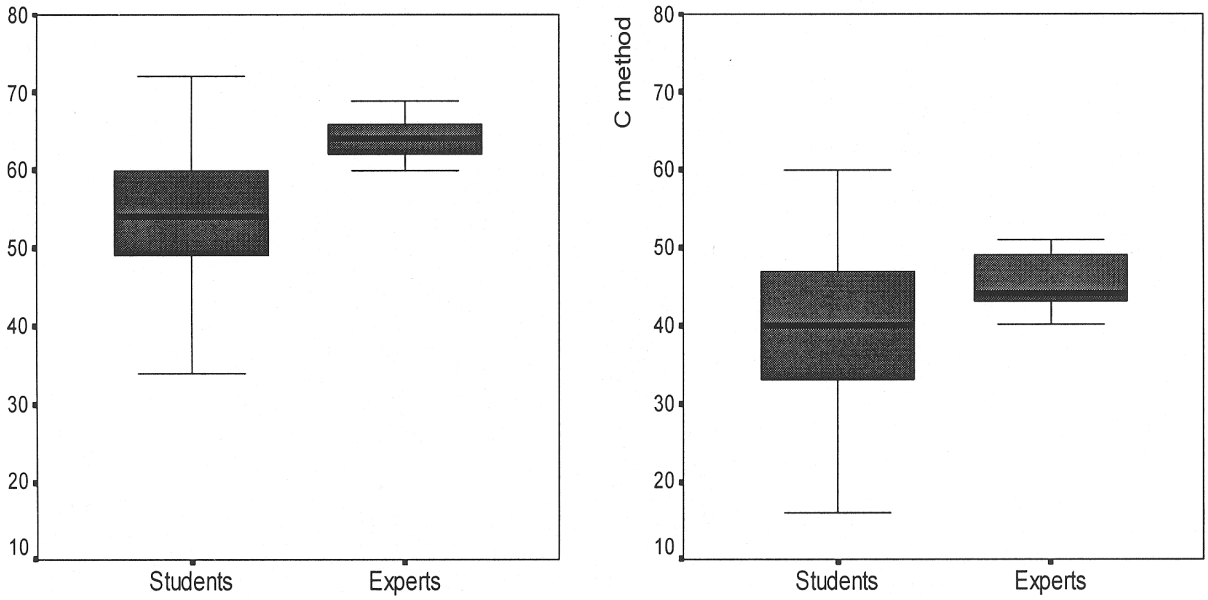

Figure 5. *Box-plot graphics with scores computed with the A method (on left) and the C method (on right).*

answers are distributed over minus 1, plus 1 (modal answer), and plus 2, whereas in the consensus condition the common answer is plus 2. An explanation for this might be that the context for the tasks differs radically in the two conditions. When an expert is alone (as is usual in practice), he or she uses only the data provided by the case, whereas within a group, the context is significantly modified by interactions with the other experts. Another argument might be that expertise lies in subtle differences in reasoning processes that disappear when experts are required to talk with each other to reach a consensus.

From a psychometric perspective, the scoring methods present some differences. The C method produces very low scores, a wide range of answers, and more variability. The A method provides higher scores and a higher mean because partial credits are given to several answers with this method, whereas with the C method only one single best answer gives credit.

The Cronbach alpha coefficient obtained with the C method (.63) was higher than the one obtained with the A method (.52). This is a consequence of the optimization of the number of items made with the C method. These coefficient values were moderate, but they were obtained on a test that had only 45 items. These values would improve with longer tests with similar item characteristics.

The A method allows a better discrimination of tested experts and students. This method is more sensible for detecting expertise than the method that follows the usual practice of asking experts to provide a consensus on each answer. When experts are judged on a consensus built by other experts, their own expertise is difficult to detect. These findings suggest that the A method has a better construct validity than method C.

An interesting observation was the production of a different ranking of respondents. This is disturbing because with the method commonly used in most certification examination, the consensus method, some students who succeeded would have failed with an aggregate scoring method that seems to have more construct validity.

It is well known that humans combine probabilities badly in their mental tasks.[19] The Baysian model of clinical reasoning is not a descriptive model of the real clinical reasoning process. It is a prescription of how clinicians should behave if they used probabilities adequately. The Smith model of categorization[15,20] (diagnosis) is descriptive. It proposes a theory of how clinician knowledge might be organized to successfully address diagnostic clinical tasks. It suggests that knowledge is organized in networks that contain expectations about values that are acceptable and not acceptable for each attribute of a clinical entity. The script concordance test is built around this assumption. The Likert scale of the test does not probe judgment about probabilities. It replicates the kinds of judgment clinicians made in their reasoning process in respect to Smith's model.

One of the major criticisms of aggregate scoring is that when it is applied blindly, wrong answers occasionally will be given credit. Experts make divergent interpretations of data in ill-defined situations met in clinical encounters. The current study confirms that. We kept all answers as valid because we considered that in a case-based rich context, each person brings his or her own experience to the situation and make his or her own interpretation. Experts differ in the multiple decisions that are made in a clinical reasoning process, whereas they usually converge toward a similar outcome. In the study we considered that if an expert makes an interpretation, a student who makes the same interpretation should receive a partial credit. That position might be challenged. An alternative method would ask criterion experts what options should be considered correct. The aggregate weights then could be applied only to these options. Deciding which method is best will be the object of further research.

In this study, we set the weights of the most popular options (modal answer) to one. This ensured that all items contributed equally to the score (at least nominally). Another method would have been to allow the weights to vary naturally depending on the level of agreement among experts. This would give greater weight to those items where there was strong consensus among the experts and a lower weight where there was disagreement. We chose the first method because students have difficulties accepting that items might have a different maximum credit while they have no way of knowing which items provide more credit.

The study has several limitations. (a) Only three clinical situations were tested, and (b) only seven experts were tested. (c) Correlation between the two methods was not very high, which may be attributable in part to the moderate reliability coefficients of both methods. (d) There was a 1-year delay between the time that criterion group experts completed the test, once alone and once in-group. We don't know if the delay was responsible for the change in experts' answers or if the context was responsible, as we suggested previously. These limitations will be addressed in subsequent research projects. Nevertheless, the study indicates that the method which is closer to the reality of clinical practice is richer and more productive to discriminate participants across their level of expertise. The consensus method seems to reduce information and validity.

This study confirms a fact found in research on clinical reasoning: In authentic tasks, experts vary substantially on their progression toward the solution. Moreover, their actual performance (aggregate method) differs from what they expect from students (consensus method). This suggests that, in assessment endeavors, the usual practice of asking experts to reach a con-

sensus on each item reduces authenticity and hinders the detection of expertise. If these results are confirmed by other researches, this practice should be reconsidered, because the aggregate method appears better for assessing complex performance in ill-defined situations.

References

1. Grant J, Marsden P. Primary knowledge, medical education and consultant expertise. *Medical Education* 1988;22:173–9.
2. Barrows HS, Feightner JW, Neufeld VR, Norman GR. *Analysis of the clinical methods of medical students and physicians*. Final Report to the Province of Ontario Department of Health, 1978.
3. Elstein AS, Shulman LS, Sprafka SA. *Medical problem solving: An analysis of clinical reasoning*. Cambridge, MA: Harvard University Press, 1978.
4. Kassirer JP. Teaching clinical medicine by iterative hypothesis testing: Let's preach what we practice. *New England Journal of Medicine* 1983;309:921–3.
5. Chamberland M, Des Marchais JE, Charlin B. Carrying PBL into the clerkship: A second reform in the Sherbrooke curriculum. *Annals of Community-Oriented Education* 1992;5:235–47.
6. Barrows HS. *What your tutor may never tell you: A guide for medical students in problem-based learning* (rev. ed.). Springfield, IL: Southern Illinois University School of Medicine, 1996.
7. Swanson DB, Norcini JJ, Grosso LJ. Assessment of clinical competence: Written and computer-based simulations. *Assessment and Evaluation in Higher Education* 1987;12:220–46.
8. Sedlacek WE, Nattress LW. A technique for determining the validity of patient management problems. *Journal of Medical Education* 1972;47:263–6.
9. Mazzuca SA, Cohen SJ. Scoring patient management problems: External validation of expert consensus. *Evaluation of Health Professionals* 1982;5:210–7.
10. Norman GR. Objective measurement of clinical performance. *Medical Education* 1985;19:43–7.
11. Norcini JJ, Shea JA, Day SC. The use of the aggregate scoring for a recertification examination. *Evaluation and the Health Professions* 1990;13:241–51.
12. Charlin B, Roy L, Brailovsky CA, Van der Vleuten CPM. The Script Concordance Test: A tool to assess the reflective clinician. *Teaching and Learning in Medical Education* 2000;12:189–95.
13. Feltovich PJ, Barrows HS. Issues of generality in medical problem solving. In HG Schmidt, ML De Volder (Eds.), *Tutorials in problem-based learning: A new direction in teaching the health professions* (pp. 128–142). Assen, Holland: Van Gorcum, 1984.
14. Schmidt HG, Norman GR, Boshuizen HPA. A cognitive perspective on medical expertise: Theory and implications. *Academic Medicine* 1990; 65:611–21.
15. Charlin B, Tardif J, Boshuizen HPA. Scripts and medical diagnostic knowledge: Theory and applications for clinical reasoning instruction and research. *Academic Medicine* 2000;75:182–90.
16. Charlin B, Brailovsky CA, Brazeau-Lamontagne L, Samson L, Leduc C. Script questionnaires: Their use for assessment of diagnostic knowledge in radiology. *Medical Teacher* 1998;20:567–71.
17. Charlin B, Brailovsky CA, Leduc C, Blouin D. The Diagnostic Script Questionnaire: A new tool to assess a specific dimension of clinical competence. *Advances in Heath Sciences Education* 1998;3:51–8.
18. Brailovsky C, Charlin B, Beausoleil S, Coté S, van der Vleuten C. Measurement of clinical reflective capacity early in training as a predictor of clinical reasoning performance at the end of residency: An exploratory study on the Script Concordance Test. *Medical Education* 2001;35:430–6.
19. Kahneman D, Slovic P, Tversky A (Eds.). *Judgement under uncertainty: Heuristics and biases*. Cambridge: Cambridge University Press, 1982.
20. Smith EE. Concepts and induction. In MI Posner (Ed.), *Foundations of cognitive science* (pp. 501–526). Cambridge, MA: MIT Press, 1989.

Received 27 June 2001
Final revision received 5 December 2001