The American Journal of Surgery 193 (2007) 248–251

Association for Surgical Education

Is the script concordance test a valid instrument for assessment of intraoperative decision-making skills?

Sarkis Meterissian, M.D.[a,b,c,*], Brent Zabolotny, M.D.[a,b], Robert Gagnon, M.Sc.[d], Bernard Charlin, M.D., Ph.D.[d]

[a]*Royal Victoria Hospital, 687 Pine Ave. W. Sl0.22, Montreal, Quebec H3A 1A1, Canada*
[b]*Division of General Surgery, McGill University, Canada*
[c]*Center for Medical Education, McGill University, Canada*
[d]*Unit for Research and Development in Health Sciences Education, University of Montreal, Montreal, Canada*

Manuscript received April 5, 2006; revised manuscript October 2, 2006

Presented at the 2006 meeting of the Association for Surgical Education, Tucson, AZ, March 23–26, 2006

Abstract

Background: Intraoperative decision making requires both knowledge and experience. The script concordance test (SCT), based on cognitive psychology script theory, is a new tool of clinical-reasoning assessment that may be used to evaluate a candidate's approach to ill-defined problems encountered in the operating room.

Methods: To develop and validate an SCT for assessment of intraoperative decision making. One hundred questions were prepared based on the objectives for residency training of the American Board of Surgery. These questions were reviewed for face and content validity by 3 board-certified general surgeons. The SCT was administered to 36 general surgical residents ranging from R1 to R5. The scoring grid was obtained by giving the test to 10 board-certified general surgeons who completed the test independently. Aggregate scoring was used. After question optimization, the final test used for statistical analysis was composed of 62 questions.

Results: The test had excellent reliability (Cronbach α, .85). Scores increased with higher levels of training except for a small decrease in the R5 scores (R1, 52.5 ± 9.9; R2, 62.4 ± 5.1; R3, 68.3 ± 9.2; R4, 75.7 ± 9.6; R5, 68 ± 6.4) There was a significant difference in scores between the junior (R1 + R2) and senior (R3 + R4 + R5) residents: 56.8 ± 9.5 versus 70.2 ± 8.8 ($P < .0001$).

Conclusions: SCT, applied to the assessment of intraoperative clinical judgment, can discriminate successfully between junior and senior residents. Results from an SCT test must be compared with the present gold standard, the oral examination, to better determine its place as an assessment tool. © 2007 Excerpta Medica Inc. All rights reserved.

Keywords: Clinical reasoning; General surgery; Script concordance; Decision making

Clinical medicine is fraught with ill-defined problems that clinicians solve in a myriad of ways [1]. Experienced practitioners possess elaborate networks of knowledge fitted to the regular tasks they perform called *scripts* [2,3]. These scripts allow the clinician to determine the diagnosis, strategies of investigation, or treatment options. Scripts begin to appear during medical school and are refined over years of clinical experience [4]. Problem solving in the operating room requires a mixture of knowledge and experience. Problems are encountered that can force the surgeon to deviate from his/her preoperative surgical plan and such decisions under pressure can significantly affect the patient's outcome. The script concordance approach, which is based on cognitive psychology script theory [3], may provide a way to build a theory-based tool to assess decision-making skills such as those in the intraoperative setting. The script concordance test (SCT) is a new tool of clinical reasoning assessment developed by Charlin et al [5] based on theoretic and empiric findings about clinical reasoning

Winner of the Haemonetics Most Outstanding Paper Award.

* Corresponding author. Tel.: +1-514-934-1934, ext. 36631; fax: +1-514-843-1503.

E-mail address: sarkis.meterissian@muhc.mcgill.ca

doi:10.1016/j.amjsurg.2006.10.012

[2,4]. Unlike simple multiple-choice questions, the SCT assesses the reasoning skills of examinees and how they actively process information to confirm or eliminate hypotheses with a series of qualitative judgments. The purpose of this study was to develop and validate an SCT as a measure of intraoperative decision-making skills in a general surgery residency program.

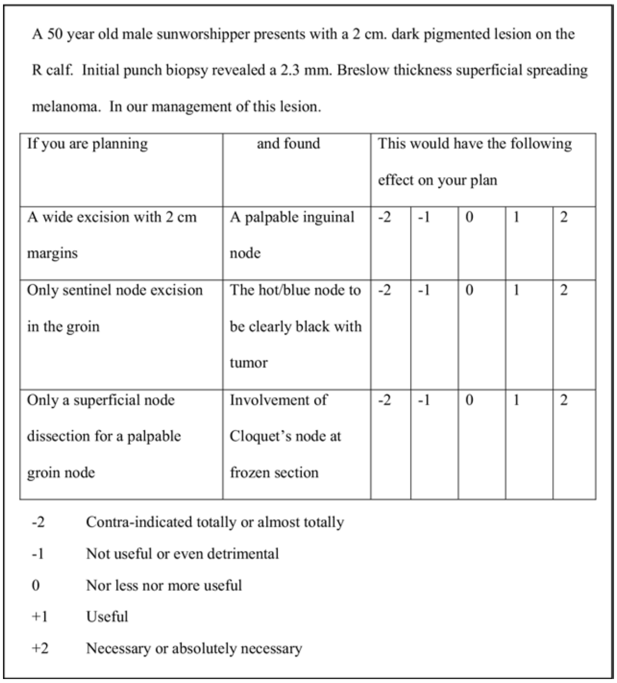
Methods

Construction of the test

Each SCT item requires a stem and between 2 to 3 questions. Based on Charlin's work [3], 50 to 60 questions are needed in an SCT to achieve a reliability coefficient (Cronbach α) of .80. Each item of our SCT examination was built so that reflection in action would be necessary to answer it. When preparing the clinical vignette an attempt was made to keep the clinical scenario authentic but to require reasoning skills and some experience. A total of 196 questions initially were developed by the authors (S.M. and B.Z.) (Fig. 1). Each question had an answer key in the form of a 5-point Likert scale (-2, -1, 0, $+1$, $+2$).

Face and content validity

The initial 196-question examination was assessed by 3 staff general surgeons for face validity (whether the question actually addressed a realistic intraoperative dilemma and whether it tested decision-making skills) and content validity (whether the examination addressed the objectives of training of both the Royal College of Surgeons of Canada and the American Board of Surgery). This process allowed us to retain the best 100 questions for the test administered to residents.



A 50 year old male sunworshipper presents with a 2 cm. dark pigmented lesion on the R calf. Initial punch biopsy revealed a 2.3 mm. Breslow thickness superficial spreading melanoma. In our management of this lesion.

If you are planning	and found	This would have the following effect on your plan				
A wide excision with 2 cm margins	A palpable inguinal node	-2	-1	0	1	2
Only sentinel node excision in the groin	The hot/blue node to be clearly black with tumor	-2	-1	0	1	2
Only a superficial node dissection for a palpable groin node	Involvement of Cloquet's node at frozen section	-2	-1	0	1	2

-2	Contra-indicated totally or almost totally
-1	Not useful or even detrimental
0	Nor less nor more useful
+1	Useful
+2	Necessary or absolutely necessary

Fig. 1. An example of an SCT question.

Scoring

One of the unique features of the SCT is the scoring process. This scoring system takes into account the range of potential answers and allows for the variability in clinical reasoning that experts show when confronted with complex questions. There is no right or wrong answer, every choice selected by an expert receives credit. A multiple-choice examination assesses knowledge, there is a clear-cut right answer, and no variability. The oral examination is the gold standard for assessing clinical reasoning and yet it is more subjective in its scoring than the SCT. The key features of the SCT—challenging authentic clinical situations, incomplete clinical data, and multiple possible correct answers—are what distinguish it from multiple-choice questions and oral examinations. If only one answer is chosen by the experts then the question behaves more like a multiple choice question and likely is too straightforward, whereas if every item of the Likert scale is chosen then the question likely is too vague or confusing and should be discarded. To develop the scoring system the examination was administered to 10 McGill attending staff general surgeons, all of whom were within 5 years of completing their residency. Scores for each question were computed from the frequencies given to each point of the Likert-type scale. Credits for each answer were transformed proportionally to get a maximum score of 1 for modal experts' choice(s) on each item; other experts' choices received partial credit. Answers not chosen by experts received zero credit. For example if on a question 8 experts out of 10 had chosen $+1$, a resident choosing $+1$ would get 1 point (8/8). If 2 experts had chosen $+2$, then a resident choosing $+2$ would receive .25 points (2/8). Choices -1, -2, and 0 would receive 0 points. The total score for the test was the sum of credits on all items.

Participants

The 2 authors of the examination, the 3 content and face validity reviewers, and the 10 experts were all different individuals. The examination was administered to 36 of the 40 general surgical residents in the McGill program (4 of the residents missed the examination as a result of vacation or out-of-town electives). There was no time limit given. The panel of experts had the following requirements: (1) they had to be general surgeons and not subspecialists, and (2) they had to be within 5 years of graduation.

Statistical analysis

By using a 1-way analysis of variance, a total sample of 33 residents would be needed to achieve an 83% power to detect differences among the means versus the alternative of equal means using an F-test with a .05000 significance level. Reliability was estimated with the Cronbach α coefficient. Optimization of the test was performed by calculating the corrected item–total item correlation for each item and then eliminating in a stepwise manner items with item–total item correlation of less than .10. The process of optimization was stopped when no more items showed correlation of less than .10. This process ensures maximal internal consistency of the final scale (Cronbach α). Construct validity (expressed as higher concordance scores with

Table 1
Examination scores (62-item test) by resident level

Resident level	n	Mean score, %	SD, %
R1	8	52.5	9.96
R2	6	62.5	5.12
R3	9	68.3	9.19
R4	6	75.7	9.61
R5	7	68.0	6.44

higher levels of training) was tested with a 1-way analysis of variance with post hoc comparisons tests and planned contrasts. All P values at an α of less than 5% were considered significant.

Results

Examination scores

The 100-question examination was completed by all the residents within 1.5 hours. Nine residents did not answer all the questions (likely because the questions were too difficult and novel). The Cronbach α of the 100-question examination was .70. However, optimization of the examination by elimination of questions that had a correlation to the total score of less than .10 resulted in the retention of the best 62 questions with a Cronbach α of .85. Scores on the optimized examination are shown in Table 1.

Comparison of groups using the optimized 62-question test revealed a statistically significant difference between the scores of R1s and those of R3s, R4s, and R5s, respectively (Table 1). When comparing junior residents with senior residents, a statistically significant difference was noted. Junior residents (R1 + R2) had a combined score of 56.8% (SD, 9.45) versus 70.24% (SD, 8.83) for the senior residents (R3 + R4 + R5). This difference proved to be statistically significant ($P < .0001$).

Comments

The assessment of clinical reasoning is challenging partly because we are unsure of its exact development and refinement. One hypothesis of how clinicians develop decision-making skills is based on cognitive psychology script theory [3]. An assessment tool based on script theory, the SCT, may represent a novel and effective means of clinical-reasoning assessment. The script concordance approach has 3 fundamental principles [6]: (1) examinees are faced with a challenging authentic clinical situation in which several options are possible, (2) responses are given according to a Likert scale that reflects script clinical reasoning theory, and (3) scoring is based on the aggregate scoring method to take into account the variability of the clinical-reasoning process among experts.

The aggregate scoring method [7,8] is a key component of the SCT because it integrates the variability that experts show when confronted with ill-defined clinical problems into the scoring process.

In this study, a bank of 196 questions covering the field of general surgery were developed. This was an arduous task and took about 6 months. Educators are not familiar with the format of SCT questions and tend to make questions either too easy (multiple-choice type) or too difficult (answers are too spread out). Candidates also are not familiar with the format and thus find the examination too difficult at first. But, as noted here, with increasing exposure the format of the SCT became more understandable and even fun. The other stumbling block we encountered was finding 10 board-certified general surgeons who were not overly subspecialized and therefore could be used as our reference panel. To avoid this potential pitfall, we created a reference panel of surgeons who were within 5 years of graduation and therefore showed temporal proximity to their Board examinations and surgeons who had a true general surgical practice. The other option would have been to give topic-specific parts of the examination to content experts (ie, colorectal questions to colorectal surgeons). We opted against this because we believed that our approach was more akin to the present gold standard for clinical-reasoning assessment, the oral examination, in which examiners ask questions in areas outside their domain of clinical expertise.

The SCT developed in this study showed excellent internal reliability as shown by the high Cronbach α scores. In addition, test scores increased from R1s to R5s, with a small decrease in the scores of R5s. Junior residents (R1 + R2) had significantly lower scores than senior residents (R3 + R4 + R5). At present, we cannot show that senior residents have significantly greater intraoperative decision-making skills because we cannot measure this objectively. In-training evaluations are woefully subjective. But, clearly, decision-making skills should improve with increasing experience. Thus, the score on the SCT should increase with resident year (otherwise surgical residencies would only be 2–3 years in duration). In Canada, the R1 and R2 years are considered *core surgery training* and R3, R4, and R5 are considered *senior specialty training* years. All R1s and R2s are officially in the core surgery program regardless of their subspecialty residency program. Thus, the division at the R2/R3 level is based on objective criteria. Other studies also have shown the predictive validity of the SCT. Brailovsky et al [9] looked at whether scores obtained on the SCT taken at the end of the clerkship predicted performances on tests of clinical reasoning (short-answer management problems and simulated office orals). In a group of 24 students the SCT was highly correlated with the short-answer management problems and simulated office orals but not with an objective structured clinical examination, which assessed hands-on skills rather than clinical reasoning. The SCT also showed construct validity with scores increasing with clinical experience in a study of urology residents in a French and a Canadian University [10]. Interestingly, this study showed that candidates' scores depended on the cohort of experts used to derive the answer key, with higher scores obtained if experts were used from the same country. Similarly, one hypothesis for the decrease in R5 scores noted in our study could be that the intense studying these final-year residents had undertaken in preparation for their examinations actually may have increased their knowledge level relative to that of our expert panel. This hypothesis will be tested by rescoring the examination with an answer key derived by content experts (subspecialists) because it could be that the knowledge of final-year residents may approximate this cohort more closely.

This study shows that an SCT can be developed with face, content, and construct validity. Such a test may be used both as a formative and summative tool in the assessment of intraoperative decision-making skills during a general surgery residency.

References

[1] Grant J, Marsden P. Primary knowledge, medical education, and consultant expertise. Med Educ 1988;22:173–9.

[2] Feltovich PJ, Barrows HS. Issues of generality in medical problems solving. In: Schmidt HG, de Volder ML, eds. Tutorials in Problem-Based Learning: A New Direction in Teaching the Health Professions. Assen, The Netherlands: Van Garcum; 1984:128–42.

[3] Charlin B, Tardif J, Boshuizen HPA. Scripts and medical diagnostic knowledge: theory and applications for clinical reasoning instruction and research. Acad Med 2000;75:182–90.

[4] Schmidt HG, Norman GR, Boshuizen HPA. A cognitive perspective on medical expertise: theory and implications. Acad Med 1990;65:611–21.

[5] Charlin B, Brailovsky RL, van der Vleuten CPM. The script concordance test: a tool to assess the reflective clinician. Teaching and learning. Med Educ 2000;12:189–95.

[6] Charlin B, van der Vleuten CPM. Standardized assessment of reasoning in contents of uncertainty: the script-concordance approach. Eval Health Prof 2004;27:304–19.

[7] Norman GR. Objective measurement of clinical performance. Med Educ 1985;19:43–7.

[8] Norcini JJ, Shea JA, Day SC. The use of the aggregate scoring method for a recertification examination. Eval Health Prof 1990;13: 241–51.

[9] Brailovsky C, Charlin B, Beausoleil S, et al. 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. Med Educ 2001;35:430–6.

[10] Sibert L, Charlin B, Corcos J, et al. Stability of clinical reasoning assessment results with the script concordance test across two different linguistic, cultural and learning environments. Med Teacher 2002; 24:537–42.