The American Journal of Surgery (2012) 203, 530–534

The American Journal of Surgery®

Association for Surgical Education

The script concordance test as a measure of clinical reasoning: a national validation study

Thamer Nouh, M.B.B.S., F.R.C.P.C., F.R.C.S.C.[a], Marylise Boutros, B.Sc., M.D., F.R.C.S.C.[a], Robert Gagnon, M.Sc.[b], Susan Reid, M.D., F.R.C.S.C.[c], Kenneth Leslie, M.D., M.H.P.E., F.R.C.S.C.[d], David Pace, M.D., M.B.A., F.R.C.S.C.[e], Dennis Pitt, M.D., F.R.C.S.C.[f], Ross Walker, M.D., F.R.C.S.C., F.A.C.S.[g], Daniel Schiller, M.D., F.R.C.S.C., M.Sc.[h], Anthony MacLean, M.D., F.R.C.S.C., F.A.C.S.[i], Morad Hameed, M.D., M.P.H., F.R.C.S.C., F.A.C.S.[j], Paola Fata, M.D., F.R.C.S.C.[a], Bernard Charlin, M.D., Ph.D.[b], Sarkis H. Meterissian, M.D.C.M., M.Sc., F.R.C.S.C., F.A.C.S.[a,k,*]

[a]Department of Surgery, McGill University, Montreal, Quebec, Canada; [b]Faculty of Medicine, University of Montréal, Montreal, Quebec, Canada; [c]Department of Surgery, McMaster University, Hamilton, Ontario, Canada; [d]Department of Surgery, University of Western Ontario, London, Ontario, Canada; [e]Department of Surgery, Memorial University, St. John's, Newfoundland; [f]Department of Surgery, University of Ottawa, Ottawa, Ontario, Canada; [g]Department of Surgery, Queen's University, Kingston, Ontario, Canada; [h]Department of Surgery, University of Alberta, Edmonton, Alberta, Canada; [i]Department of Surgery, University of Calgary, Calgary, Alberta, Canada; [j]Department of Surgery, University of British Columbia, Vancouver, British Columbia, Canada; [k]Center for Medical Education, McGill University, Montreal, Quebec, Canada.

Abstract

INTRODUCTION: The script concordance test (SCT) is an innovative tool for clinical reasoning assessment. It has previously been shown to be a reliable and valid measure of clinical reasoning among general surgical residents.

PURPOSE: To determine if the SCT maintained its validity and reliability when administered on a national level.

METHODS: The test was administered to 202 residents (51 R1, 45 R2, 45 R3, 28 R4, and 33 R5) in 9 general surgery programs across Canada.

RESULTS: The optimized version of the test had a reliability (Cronbach alpha) of .85. Scores increased progressively from R1 (64.5 ± 7.6) to R2 (69.5 ± 5.8) to R3 (69.9 ± 6.7) to R4 (72.0 ± 6.2) with a dip in the R5s (68.3 ± 8.6). The test was able to differentiate junior (R1+ R2 = 66.8 ± 7.2) from senior residents (R3 + R4 + R5 = 70.0 ± 7.3, $P = .001$) across all the programs.

Presented at the 2011 Meeting of the Association for Surgical Education.

* Corresponding author. Tel.: +1-514-934-1934; fax: +1-514-843-1454.
E-mail address: sarkis.meterissian@muhc.mcgill.ca

Manuscript received April 18, 2011; revised manuscript November 22, 2011

0002-9610/$ - see front matter
doi:10.1016/j.amjsurg.2011.11.006

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

CONCLUSIONS: The SCT maintained its reliability and validity as a measure of intraoperative clinical reasoning among general surgical residents when administered across multiple centers. We believe that the SCT can be developed to measure clinical reasoning in high-stakes national examinations.

Training of general surgeons requires the acquisition of knowledge, technical skills, and experience. This is currently accomplished through residency training programs. These programs provide trainees with a structured curriculum to acquire knowledge, technical skill training through the exposure to the operating room, and the chance to build their experience in a supervised environment. Physicians who successfully complete their training become eligible for credentialing examinations. The assessment of these requirements is performed via a combination of subjective (ie, in-training evaluations) and objective assessment tools. These objective examinations are composed of both multiple-choice and oral-type questions. The multiple-choice questions are generally accepted as a valid and reliable measure of knowledge acquisition. By contrast, oral-type questions test knowledge, clinical reasoning, and decision-making skills.[1] Although oral examinations can assess the clinical reasoning and decision-making skills needed to solve ill-defined problems, they are limited by difficulties in proper standardization, reliable scoring, and administration to a large group of examinees. With cognitive psychology now forming the major conceptual framework in medical education,[2] it is necessary to develop a reliable and valid method of assessing clinical reasoning that measures its process as well as its outcome.[3]

The script concordance test (SCT) is a tool of clinical reasoning assessment that is based on cognitive psychology script theory.[4] The theory proposes that when physicians are faced with clinical problems, they mobilize sets of knowledge (their scripts) to understand the situation and come to clinical decisions.[4,5] These scripts are used daily in clinical practice and are refined with experience.[6]

The SCT has been shown to be a reliable measure of the examinees' skills in using their knowledge to confirm or eliminate a clinical hypothesis in relation to ill-defined problems across multiple medical specialties[6–10] including intraoperative problems encountered by surgeons.[11,12] We have previously shown, in a single-institution study, that the SCT is reliable and valid in differentiating between the intraoperative clinical reasoning skills of junior and senior general surgical residents.[11] The purpose of this study was to determine if the SCT maintained its validity and reliability in differentiating between the intraoperative clinical reasoning skills of junior and senior residents when administered across 9 Canadian general surgery programs, which is a natural next step in the assessment of this innovative tool for its possible use in high-stakes examinations.

Methods

The examination candidates

Of the 16 general surgery programs across Canada, 9 program directors agreed to participate in this study. The examination was administered to 202 general surgical residents enrolled across 9 Canadian universities: Alberta, British Columbia, Calgary, McGill, McMaster, Memorial, Ottawa, Queen's, and Western Ontario. The resident's level of training ranged between postgraduate training years 1 through 5 (R1–R5).

For analysis, the residents were also divided into 2 larger groups: junior and senior. This division follows the Royal College of Canada's "specialty training requirements in general surgery."[13] According to these requirements, the initial period of postgraduate training (R1 and R2) is considered junior years and is required to acquire the knowledge, skills, and attitudes underlying the basics to the practice of surgery in general.[14] In the remaining period of postgraduate training (R3, R4, and R5), residents assume a more senior role; they are responsible for more advanced aspects of patient care, participate more in decision making, and supervise the junior residents.

Development of the test

The SCT administered to the residents consisted of new questions that were specifically prepared for this study following the same steps described in our previous publication.[11] Four of the 9 program directors (SM, DP, SR, and KL) participated in writing the new examination items. Each SCT item consisted of a scenario and between 2 and 6 questions that covered issues inherent to the scenario.[12] Each item of our SCT examination was constructed so that reflection in action would be necessary to answer the questions. When preparing the clinical vignette, an attempt was made to keep it authentic but to require reasoning skills and some clinical experience. Each question had an answer key in the form of a 5-point Likert scale (-2, -1, 0, $+1$, and $+2$), ranging from completely contraindicated (-2) to completely indicated ($+2$).

According to Fournier et al,[15] an SCT should have 20 items with 60 questions for each hour of testing to achieve a reliability coefficient (Cronbach α) higher than 0.75. It has been shown that having 3 questions for each scenario achieved the best reliability because the addition of more questions to a scenario led to a minimal increase in reliability when compared with the addition of more scenarios.[16]

Four of the authors (board-certified general surgeons) assessed the initial SCT question database containing over 300 questions for whether each question actually addressed a realistic intraoperative dilemma and if it tested decision-making skills. This process allowed us to retain the best 153 questions for the eventual test. All the retained questions addressed an objective of training of both the Royal College of Surgeons of Canada and the American Board of Surgery.

Scoring

Scoring takes into account the range of potential answers and allows for the variability in clinical reasoning that experts show when confronted with complex questions. Every choice selected by an expert received credit. To develop the scoring grid, the examination was administered to 22 general surgeons from the 9 participating universities, all of whom volunteered to participate as experts. Scores for each question were computed from the answers chosen by all the experts. Credit for each answer was transformed proportionally to get a maximum score of 1 for the modal experts' choice on each item; other experts' choices for that question received partial credit. Choices not selected by any expert received 0 credit. For example, if on a question 17 experts out of 22 had chosen +1, a resident choosing +1 would get 1 point (17/17). If 5 experts had chosen +2, then a resident choosing +2 would receive .29 points (5/17). Choices −1, −2, and 0 would receive 0 points. The total score for the test was the sum of the credits on all items.

Statistical analysis

Reliability was estimated using the Cronbach α coefficient. The test was optimized by calculating the corrected item/total item correlation for each question and iteratively eliminating questions with a negative correlation. The score used was the sum of scores on retained questions, and no scenario-based analysis was performed. The process of optimization was stopped when no more questions showed a negative correlation. This process ensured maximal internal consistency of the final examination (Cronbach α). The relationship between the final SCT score (representing the level of concordance between the residence and the experts) and the level of training (R level and junior/senior level) was tested with a 1-way analysis of variance. As stated earlier, junior (R1 and R2) and senior (R3, R4, and R5) residents are expected to be qualitatively different in terms of decision-making skills. To compare the variability of scores between groups, a variability coefficient was calculated (standard deviation divided by the mean and multiplied by 100). All P values at an α of less than 5% were considered significant.

Table 1 SCT mean score (131 items) by resident level

Resident level	n	Mean score, %	SD, %	Variability coefficient
R1	51	64.5	7.6	11,8
R2	45	69.5	5.8	8,3
R3	45	69.9	6.7	9,6
R4	28	72.0	6.2	8,6
R5	33	68.3	8.6	12,6

SD = standard deviation.

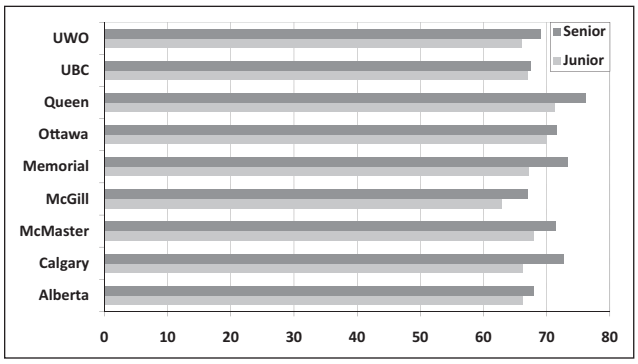

Figure 1 Junior and senior mean scores (131 items) by university. UBC, British Columbia; UWO, Western Ontario.

Results

All 202 residents completed the 153-question SCT within the 3-hour examination period. Optimization of the examination through the elimination of items with a negative item/total item correlation resulted in the retention of 131 items. The final optimized examination had a Cronbach α of .85 compared with a Cronbach α of .81 before optimization.

Scores increased progressively from R1 to R4 with a dip in the scores of the R5s ($F_4 = 6.2$, $P < .001$, Table 1). Across the 9 general surgery programs, 96 junior residents (R1 and R2) wrote the test compared with 106 senior residents (R3–R5). The test was able to differentiate junior (R1 + R2 = 66.8 ± 7.2) from senior residents (R3 + R4 + R5 = 70.0 ± 7.3, $F_{1,200} = 10.8$, $P < .001$, Fig. 1).

When assessing how junior residents performed compared with seniors within different sites of training, the SCT was able to differentiate junior from senior residents (Table 2). The site of training had a significant effect on scores ($F_{8,184} = 3.0$, $P = .003$), but there was no interaction between the site of training and the level of training ($F_{8,184} = .4$, $P = .92$). This indicates that the difference between the junior and senior groups was similar across sites.

Comments

The primary objective of this study was to assess the SCT's ability to distinguish between the intraoperative decision-making skills of junior and senior general surgical

Table 2 Junior and senior mean scores (standard deviation) by university

Site	Junior residents	n	Senior residents	n
Alberta	66.3 (7.6)	21	68.0 (5.8)	15
Calgary	66.2 (8.7)	9	72.8 (6.5)	8
McMaster	68.0 (7.3)	9	71.4 (7.4)	16
McGill	63.0 (7.2)	15	67.0 (7.8)	24
Memorial	67.2 (1.6)	5	73.4 (4.0)	5
Ottawa	70.0 (6.4)	10	71.7 (7.4)	9
Queen	71.3 (3.9)	8	76.2 (3.0)	9
UBC	67.0 (7.6)	10	67.5 (6.7)	10
Western Ontario	66.1 (9.3)	9	69.1 (8.9)	10
Total	66.8 (7.2)	96	70.0 (7.3)	106

UBC = British Columbia.

residents on a national scale. In our previous study, we showed that the SCT differentiated between junior and senior residents in a single-university cohort of general surgical residents.[11] In this study, we show that the SCT is a reliable and valid measure of distinguishing between the intraoperative decision-making skills of junior and senior general surgical residents when used across the 9 Canadian general surgery programs. Although the optimized test maintained a high internal reliability and was able to differentiate between junior and senior residents, we did not observe a significant drop in variability from R1 to R5. In fact, there is a drop in variability between R1 and R2 (coefficient of variability 11.8 and 8.3); a similar variability among R2, R3, and R4; and a rise with R5 (12.6). These observations are rather unusual; in most SCT studies, the variability of scores drops with expertise level. The analysis of scores from the sites suggests that this phenomenon is not constant; in 3 sites, noticeable drops in variability are observed, whereas other patterns can be seen in other sites. In light of the present experience and literature with SCT testing, no clear interpretation of this phenomenon can be proposed.

As in our previous study,[11] we documented an increase in scores with resident progression from R1 to R4 with the same decrease in the scores of the R5s. We attempted to explain this drop in the scores of the R5s using 2 hypotheses. In our first hypothesis, we assumed that this drop could be a result of variability between how the R5s understand and use the Likert scale (the difference between what a +2 means compared with a +1) as opposed to the expert panel. This variation in understanding has been proposed in the literature as a cause for the lack of concordance between experts and candidates.[17] To eliminate this as a potential reason for the drop in the scores of the R5s, we rescored the test after converting the 5-point Likert scale (−2, −1, 0, +1, and +2) to a 3-point Likert scale (−, 0, and +). The test still maintained its internal reliability (Cronbach α = .81), but again it showed progressively higher scores as residents progressed from R1 to R4 and the same decrease in the scores of the R5s. Although the conversion did not resolve the drop in the scores of the R5s, its effect on the internal reliability and construct validity is consistent with the findings of Bland et al.[18] In their article, the authors compared 5 different scoring systems for SCT assessment, including the aggregate 5-point and 3-point Likert scales used in our study. They found a strong absolute correlation and relationship between the scores calculated when using both scales.[18] They argued that the 5-point Likert scale added little to the SCT in terms of reliability and validity. Based on their findings and our results, it would appear that using an aggregate 3-point Likert scale could be simpler and has little effect on the performance of the test. However, adopting the 3-point Likert scale risks losing information relating to the degree of confidence examinees have in their response; it potentially eliminates all variability present in day-to-day clinical practice, and it risks transforming the questions into multiple-choice best answer–type questions.

Our second hypothesis is based on an inherent feature of the SCT. It has been shown in the literature that candidates' scores vary depending on the cohort of experts used to develop the scoring grid.[19] We hypothesize that this drop might be explained by the fact that in their preparation for their board examination the R5s may have increased their knowledge level relative to that of the expert panel, and, hence, they may be answering at a level of a subspecialist. A study is currently ongoing to explore the validity of this hypothesis by developing a scoring grid to each group of subspecialty questions that will be based on the answers of experts of that subspecialty. This important study will determine whether, in the future, SCTs are scored by experts taking the entire examination or by subspecialists taking only parts of the examination relevant to their specialty.

Although studies in many specialties have shown the SCT to be both reliable and valid, they have included a relatively small set of items and candidates.[3,7–9,11,12,20] Our study has the largest number of participants in the literature including 202 residents enrolled in general surgery training across 9 universities. Administering and scoring this examination to a large number of candidates across multiple sites can be labor intensive. It is very useful that the SCT lends itself very well to being administered as an online examination.[3,7] This not only provides an easy scoring method but also allows the inclusion of multimedia (ie, pictures, audio, and video) to clinical scenarios and to the information provided in the questions further expanding the test's ability to examine the candidate's decision-making skills and reasoning by providing a scenario that is closer to actual clinical practice. Other limitations to the widespread application of the SCT are the difficulty in preparing items that test clinical reasoning rather than knowledge and the difficulty in recruiting experts for examination scoring. Despite these limitations, we believe that the SCT is one of the few available approaches to clinical reasoning assessment at the present time. The only other approach that appears to be gaining in popularity is the combination of think-aloud and concept mapping. Pottier et al[21] used this new method to

identify the style of clinical reasoning in medical students and experts. They showed that the combination of think-aloud and concept mapping protocols could reliably assess the trainees' approach to problem solving. In contrast, the SCT does not analyze the problem-solving approach of trainees but rather determines whether the end result matches that of experts and therein lies the unique value of using the SCT as a measure of clinical reasoning proficiency.

There is a great need for a measure of decision making and clinical reasoning in surgical residency training programs and in certifying examinations. We believe, given the results of this study, that the SCT can be developed to address this need. Piloting the test on a national basis by general surgical associations would allow us to examine the performance of surgical residents (R1s–R5s) compared with surgeons in general practice and surgeons in specialized practice. This could help explain and resolve issues like the dip in the R5 scores as a step toward getting the test ready to be used as a measure of decision making and clinical reasoning within residency training programs and potentially in high-stakes national examinations.

References

1. Schön DA. The Reflective Practitioner: How Professionals Think in Action. New York, NY: Basic Books; 1983.
2. Irby DM. Shifting paradigms of research in medical education. Acad Med 1990;65:622–3.
3. Sibert L, Darmoni SJ, Dahamna B, et al. Online clinical reasoning assessment with the script concordance test: a feasibility study. BMC Med Inform Decis Mak 2005;5:18.
4. Charlin B, Roy L, Brailovsky C, et al. The script concordance test: a tool to assess the reflective clinician. Teach Learn Med 2000;12:189–95.
5. Charlin B, Tardif J, Boshuizen HP. Scripts and medical diagnostic knowledge: theory and applications for clinical reasoning instruction and research. Acad Med 2000;75:182–90.
6. Meterissian SH. A novel method of assessing clinical reasoning in surgical residents. Surg Innov 2006;13:115–9.
7. Sibert L, Darmoni SJ, Dahamna B, et al. On line clinical reasoning assessment with script concordance test in urology: results of a French pilot study. BMC Med Educ 2006;6:45.
8. Carrière B, Gagnon R, Charlin B, et al. Assessing clinical reasoning in pediatric emergency medicine: validity evidence for a script concordance test. Ann Emerg Med 2009;53:647–52.
9. Lambert C, Gagnon R, Nguyen D, et al. The script concordance test in radiation oncology: validation study of a new tool to assess clinical reasoning. Radiol Oncol 2009;4:7.
10. Lubarsky S, Chalk C, Kazitani D, et al. The script concordance test: a new tool assessing clinical judgement in neurology. Can J Neurol Sci 2009;36:326–31.
11. Meterissian S, Zabolotny B, Gagnon R, et al. Is the script concordance test a valid instrument for assessment of intraoperative decision-making skills? Am J Surg 2007;193:248–51.
12. Park AJ, Barber MD, Bent AE, et al. Assessment of intraoperative judgment during gynecologic surgery using the script concordance test. Am J Obstet Gynecol 2010;203:e1–6.
13. Specialty training requirements in general surgery. The Royal College of Physicians and Surgeons of Canada; 2010. Available at: http://rcpsc.medical.org/residency/certification/training/gen_surg_e.pdf. Accessed June 11, 2011.
14. Objectives of Surgical Foundations Training. The Royal College of Physicians and Surgeons of Canada; 2010. Available at: http://rcpsc.medical.org/residency/certification/objectives/surgical_foundations_otr_e.pdf. Accessed June 11, 2011.
15. Fournier JP, Demeester A, Charlin B. Script concordance tests: guidelines for construction. BMC Med Inform Decis Mak 2008;8:18.
16. Gagnon R, Charlin B, Lambert C, et al. Script concordance testing: more cases or more questions? Adv Health Sci Educ Theory Pract 2009;14:367–75.
17. Charlin B, Desaulniers M, Gagnon R, et al. Comparison of an aggregate scoring method with a consensus scoring method in a measure of clinical reasoning capacity. Teach Learn Med 2002;14:150–6.
18. Bland AC, Kreiter CD, Gordon JA. The psychometric properties of five scoring methods applied to the script-concordance test. Acad Med 2005;80:395–9.
19. Charlin B, Tardif J, Boshuizen HP. Scripts and medical diagnostic knowledge: theory and applications for clinical reasoning instruction and research. Acad Med 2000;75:182–90.
20. Charlin B, Gagnon R, Lubarsky S, et al. Assessment in the context of uncertainty using the script concordance test: more meaning for scores. Teach Learn Med 2010;22:180–6.
21. Pottier P, Hardouin JB, Hodges BD, et al. Exploring how students think: a new method combining think-aloud and concept mapping protocols. Med Educ 2010;44:926–35.