The American Journal of Surgery (2013) 205, 137-140

The American
Journal of Surgery®

Association for Surgical Education

# Assessing clinical judgment using the Script Concordance test: the importance of using specialty-specific experts to develop the scoring key

**Andrea M. Petrucci, M.D.[a], Thamer Nouh, M.D.[b], Marylise Boutros, M.D.[a], Robert Gagnon, M.D.[c], Sarkis H. Meterissian, M.D.[a,d,*]**

*[a]Department of Surgery, McGill University, Pine Avenue West, Suite 10.22, Montreal, Quebec H3A 1A1, Canada; [b]Department of Surgery, College of Medicine, King Saud University, Riyadh, Saudi Arabia; [c]Faculty of Medicine, University of Montreal, Montreal, Quebec, Canada; [d]Center for Medical Education, McGill University, Montreal, Quebec, Canada*

**KEYWORDS:**
Clinical reasoning;
General surgery;
Script concordance

**Abstract**

**BACKGROUND:** The Script Concordance test (SCT) assesses clinical judgment. The purpose of this study was to determine whether a specialty-specific scoring key improves the validity of the SCT.

**METHODS:** Thirty experts from 6 general surgery disciplines answered questions pertaining to their area of expertise. We created a scoring key of 5 amalgamated expert panel members. The answers of 227 general surgery residents were analyzed.

**RESULTS:** The optimized test had a reliability level (Cronbach $\alpha$) of .81. Scores increased progressively throughout all levels of training, with R5s scoring higher than R4s (R1, 42.7 ± 7.1; R2, 47.6 ± 7.5; R3, 48.7 ± 6.7; R4, 49.8 ± 7.7; R5, 52.9 ± 9.3). The average score of juniors (R1s + R2s, 45.1 ± 7.6) was significantly lower ($P < .001$) than seniors (R3s + R4s + R5s, 50.4 ± 8.0).

**CONCLUSIONS:** We showed that specialty-specific experts must be used to develop the scoring key. This has important implications in the application of the SCT on a wider level.

Surgical training is in the midst of a paradigm shift. For generations, surgical residents took on the role of apprentices, devoting countless hours to training alongside a mentor, which resulted in abundant exposure to operative cases. This exposure translated into more opportunities for developing appropriate clinical decision-making skills both inside and outside of the operating room.[1] Currently, with reduced resident working hours and increasing resident duty demands, this exposure is limited, pushing surgical educators to develop more stringent and structured competency-based curricula for their residents.[1,2] Even in structured programs, the training of proficient and technically skilled surgeons requires experience and exposure in order to develop sharp intraoperative decision-making skills. Moreover, a formal assessment of intraoperative decision-making skills is not done because of the lack of reliable and valid assessment tools.

At present, graduating general surgery residents are certified based on written multiple-choice and oral case-based examinations, with the former focusing mostly on core knowledge and the latter evaluating the decision-making process.[3] Although these traditional, objective

Presented at the Meeting of the Association for Surgical Education, March 19–24, 2012, San Diego, CA.

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

Manuscript received April 5, 2012; revised manuscript June 15, 2012

0002-9610/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.amjsurg.2012.09.002

methods of assessments are reliable,[4] capturing the acquisition of good intraoperative decision-making skills may be overlooked. The Script Concordance test (SCT) has the potential to address this issue.[5] It has been validated in many fields of medicine[6,7] and surgery[7–10] including general surgery.[9] Our group created and validated an SCT that was aimed at testing the intraoperative clinical reasoning of general surgery residents at our institution.[9] Although the SCT showed construct validity, there was an unexpected decrease in the scores of the R5s compared with the R4s. The same results surfaced in a second study in which we tried to validate the SCT on a national level, testing general surgery residents from 9 Canadian programs.[11] To explain the “dip” in the scores of the R5s, we hypothesized that these graduating R5s, in preparing for their certification board examinations, had a broader knowledge of all areas in general surgery compared with members of the expert panel who often were subspecialists with excellent knowledge in their area of expertise but had less current knowledge of other areas in general surgery. For example, a colorectal surgeon would predictably answer the colorectal questions well but may not answer the “trauma” questions in the same way as a trauma surgeon or a fifth-year resident. The purpose of this study was to assess the validity of the SCT when using a specialty-specific scoring key and whether it would resolve the issue of the drop in the scores of the R5s that was observed in our previous work.[9,11]

## Methods

### Development of the SCT

The SCT used in this study is the same test used in our nationwide study in 2011.[11] The test was developed by 4 program directors following the same guidelines as previously described.[3] The test consisted of 43 clinical scenarios involving a total of 153 questions. Questions were answered using a 5-point Likert scale (ie, −2, −1, 0, +1, and +2), ranging from completely contraindicated (−2) to completely indicated (+2).

### Scoring and creation of the specialty-specific scoring grid

To develop the specialty-specific scoring grid, we identified and recruited a total of 30 board-certified general surgeons from 6 different general surgery subspecialties. These subspecialties included colorectal surgery, endocrine surgery, hepatobiliary surgery, surgical oncology, thoracic surgery, and trauma and acute care surgery. The experts were contacted via e-mail and asked to voluntarily participate in our study. There were 5 subspecialists for each of the 6 surgical subspecialties. The subspecialists who agreed to participate were sent instructions on how to answer the SCT using the Likert-type scale. Each subspecialist was given only the portion of the 153-question examination that pertained to their subspecialty. For example, colorectal surgeons answered only colorectal questions.

Considering that each subspecialist answered a subgroup of questions, we grouped together a single subspecialist from each of the 6 surgical subspecialties to create an amalgamated expert examination covering all 153 questions. Thus, answers from 6 experts were used to create 1 expert scoring key. Therefore, with the 30 subspecialists, we created 5 expert scoring keys; each was made up of 6 subspecialists.

Answers were scored according to the modal experts' choice, which was transformed proportionally to receive a maximum score of 1. All other choices selected by an expert for that same question received partial credit. Choices not selected by any expert received zero credit. For example, if on a question 4 amalgamated experts out of the 5 had chosen −2, a resident choosing −2 would get 1 point (4/4). If 1 amalgamated expert had chosen −1, then a resident choosing −1 would receive 0.25 points (1/4). Choices +1, +2, and 0 would receive 0 points. The total score for the test was the sum of credits on all items.

### Participants

To increase our numbers, we administered the SCT to 25 general surgery residents ranging from postgraduate year 1 to postgraduate year at McGill University, Montreal, Quebec, Canada, during their academic half-day. Participation was voluntary, and informed consent was obtained from all the residents. They were given a short introductory presentation explaining how to use the 5-point Likert scale to answer the questions using a few examples. The residents were given 3 hours to complete the examination. We combined our residents' answers with those of the 202 residents from our previous study[11] to create a participant pool of 227 residents. The answers of all the residents were scored and analyzed with the new specialty-specific scoring key. The institutional ethics review board reviewed and approved this study.

### Statistical analysis

As for the analysis of our national study,[11] the residents were further divided into 2 larger groups: junior and senior. This division follows the Royal College of Canada “Specialty Training Requirements in General Surgery.”[12] These requirements suggest that the initial 2 year- period of postgraduate training (R1 and R2) is considered junior years and focuses on the acquisition of the principles of surgery including basic knowledge, skills, and attitudes required for the practice of surgery in general.[13] In the following 3 years (R3, R4, and R5), residents become seniors and are expected to take on more advanced aspects of patient care, participate more in decision making, and supervise and teach the junior residents.

Statistical analysis was done following the same methods used in our national study.[11] Reliability was estimated using the Cronbach $\alpha$ coefficient. The test was optimized by calculating the corrected item/total item correlation for each question and eliminating iteratively questions with a negative correlation. 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 $\alpha$). The score used was the sum of scores on retained questions, and no scenario-based analysis was performed. As previously described,[11] the relationship between the final SCT score (representing the level of concordance between the residents and the experts) and the level of training (R level and junior/senior level) was tested with 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 on the ground of decision-making skills. To compare the variability of scores between groups, a variability coefficient was calculated (standard deviation divided by the mean). All $P$ values at an $\alpha$ of less than 5% were considered significant.

## Results

### Examination scores

All of the 227 examination papers were scored using the new specialty-specific scoring grid. The 153-question test had a Cronbach $\alpha$ of .67. After eliminating items with a negative item–to–total item correlation, we were left with an optimized examination that consisted of 100 questions. The optimized examination had a Cronbach $\alpha$ of .81.

Residents' scores increased significantly throughout all levels of training, with the R5 residents scoring higher than all the other levels (Table 1). Although the scores for R5s were not significantly higher than R4s, there was a strong statistically significant linear trend, with scores rising with the increasing level of residency ($F_{4,222} = 11.5$, $P < .001$, Fig. 1).

Our new participant pool of 227 residents consisted of 109 junior residents and 118 senior residents. The new specialty-specific scoring key successfully differentiated between junior and senior residents. The average score of junior residents (R1s + R2s = 45.1 ± 7.6) was significantly lower than the average score of senior residents (R3s + R4s + R5s = 50.4 ± 8.0, $F_{1,225} = 25.8$, $P < .001$).

**Table 1** The SCT mean score by resident level (100 items)

| Resident level | n | Mean score | SD | Variability coefficient |
|---|---|---|---|---|
| R1 | 56 | 42.7 | 7.1 | 0.15 |
| R2 | 53 | 47.6 | 7.5 | 0.16 |
| R3 | 50 | 48.7 | 6.7 | 0.14 |
| R4 | 29 | 49.8 | 7.7 | 0.15 |
| R5 | 39 | 52.9 | 9.3 | 0.17 |
| Total | 227 | 47.8 | 8.3 | 0.17 |

SD = standard deviation.

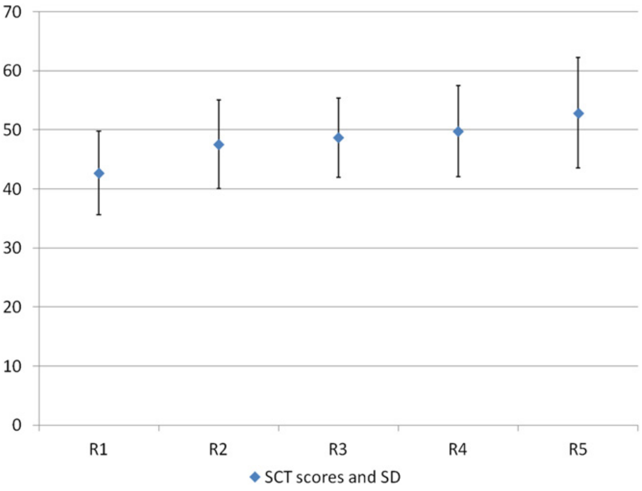

**Figure 1** The SCT mean score by resident level.

## Comments

We have shown that consideration must be given to using specialty-specific experts when creating the scoring key of an SCT. In previous single-institution and pan-Canadian studies, we showed that an SCT assessing intraoperative decision-making skills maintained its reliability and validity across all levels of general surgery residency; however, we were faced with the unexpected finding of the R5 residents scoring lower than the R4s.[8,11] Our hypothesis was that the R5s most likely have more subject-specific knowledge than the panel of experts because they are in the process of studying for their board examinations. Hence, the R5s approach situations with the knowledge of subspecialists.[11]

The rationale behind using specialty-specific experts for creating the scoring key is that general surgery is a broad specialty consisting of many different subspecialties. Each of these subspecialties encompasses an ever-expanding (one could say exploding) area of knowledge and clinical expertise. Despite this subspecialization, we expect final-year residents to have a broad array of knowledge and expertise, a trait that quickly decays after graduation and subspecialization. For instance, a surgical oncologist may not be aware of the latest evidence and current practices of trauma surgery. As a result, the scoring key used in our previous study may have been flawed because the experts were board-certified general surgeons who answered the entire examination despite lacking up-to-date expertise in certain areas. This might not have reflected an accurate representation of "expert decision making."

The development of the specialty-specific expert panel proved to be a difficult task. Gagnon et al[14] concluded that a panel of 10 is needed to achieve an acceptable reliability and a panel of 20 should ideally be used for high-stakes examinations. We needed to recruit 30 surgeons from 6 different subspecialties to come up with 5 amalgamated experts. We would have needed to recruit double that number to achieve the recommended panel of 10 or 120 experts to develop a panel of 20, which is recommended for high-stakes examinations. Despite the limitation of our panel consisting of only 5 amalgamated experts, the test achieved a reliability coefficient of .81. This compares favorably with the Cronbach $\alpha$ of the examination used in Nouh et al's study[11] of .85. A possible explanation for the high reliability coefficient of our examination may be the large number of questions used. We had 43 scenarios with 153 questions (3–4 nested questions per scenario), which is clearly higher than the optimal number of 15 to 25 scenarios containing 2 to 4 nested questions that is recommended by Gagnon et al[15] to maintain an acceptable level of reliability. Although the reliability of the examination in this study was respectable, we appreciate that the overall scores were low. This could be partially explained by the lower number of experts we had on our panel resulting in lower overall scores as described by Gagnon et al,[14] who found that mean scores varied proportionally with panel size. This is to be expected because with more experts the chances of more answers being chosen increases the potential of residents receiving partial credit on a particular question.

Published literature clearly describes various aspects of the SCT including its approach,[5] construction,[16] the number and content of scenarios and questions,[15] the expert panel,[14,17] and scoring.[18,19] In this study, we showed that consideration must also be given to using specialty-specific experts when creating the scoring key of an SCT. This is probably more important when developing an SCT to assess clinical judgment in very broad areas of medicine such as general surgery or internal medicine. The results of this study need to be verified on a national level because this would allow for the recruitment of more experts. If our results are confirmed, then the SCT can be used as a formative tool to diagnose decision-making problems during residency as well as an educational tool to help residents understand how specialty-specific experts think. The results of this latest study further extend our knowledge of the SCT and will help us determine how to use it as an educational tool in surgical residency programs.

## References

1. Flin R, Youngson G, Yule S. How do surgeons make intraoperative decisions? Qual Saf Health Care 2007;16:235–9.
2. Poulose BK, Ray WA, Arbogast PG, et al. Resident work hour limits and patient safety. Ann Surg 2005;241:847–56; discussion 56–60.
3. Meterissian SH. A novel method of assessing clinical reasoning in surgical residents. Surg Innov 2006;13:115–9.
4. Epstein RM, Hundert EM. Defining and assessing professional competence. JAMA 2002;287:226–35.
5. Charlin B, van der Vleuten C. Standardized assessment of reasoning in contexts of uncertainty: the script concordance approach. Eval Health Prof 2004;27:304–19.
6. Ruiz JG, Tunuguntla R, Charlin B, et al. The script concordance test as a measure of clinical reasoning skills in geriatric urinary incontinence. J Am Geriatr Soc 2010;58:2178–84.
7. 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 experimental study on the script concordance test. Med Educ 2001;35:430–6.
8. Sibert L, Charlin B, Corcos J, et al. Assessment of clinical reasoning competence in urology with the script concordance test: an exploratory study across two sites from different countries. Eur Urol 2002;41:227–33.
9. 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.
10. 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;203:240e1–6.
11. Nouh T, Boutros M, Gagnon R, et al. The script concordance test as a measure of clinical reasoning: a national validation study. Am J Surg 2012;203:530–4.
12. Specialty Training Requirements in General Surgery. The Royal College of Physicians and Surgeons of Canada, Ottawa, Canada; 2010.
13. Objectives of Surgical Foundations Training. The Royal College of Physicians and Surgeons of Canada, Ottawa, Canada; 2010.
14. Gagnon R, Charlin B, Coletti M, et al. Assessment in the context of uncertainty: how many members are needed on the panel of reference of a script concordance test? Med Educ 2005;39:284–91.
15. 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.
16. Fournier JP, Demeester A, Charlin B. Script concordance tests: guidelines for construction. BMC Med Inform Decis Mak 2008;8:18.
17. Charlin B, Gagnon R, Pelletier J, et al. Assessment of clinical reasoning in the context of uncertainty: the effect of variability within the reference panel. Med Educ 2006;40:848–54.
18. 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.
19. Gagnon R, Lubarsky S, Lambert C, et al. Optimization of answer keys for script concordance testing: should we exclude deviant panelists, deviant responses, or neither? Adv Health Sci Educ Theory Pract 2011;16:601–8.