EDUCATION

Assessment of intraoperative judgment during gynecologic surgery using the Script Concordance Test

Amy J. Park, MD; Matthew D. Barber, MD, MHS; Alfred E. Bent, MD; Yashika T. Dooley, MD; Christina Dancz, MD; Gary Sutkin, MD; J. Eric Jelovsek, MD

OBJECTIVE: We sought to develop a valid, reliable assessment of intraoperative judgment by residents during gynecologic surgery based on Script Concordance Theory.

STUDY DESIGN: This was a multicenter prospective study involving 5 obstetrics and gynecology residency programs. Surgeons from each site generated case scenarios based on common gynecologic procedures. Construct validity was evaluated by correlating scores to training level, in-service examinations, and surgical skill and experience using a Global Rating Scale of Operative Performance and case volumes.

RESULTS: A final test that included 42 case scenarios was administered to 75 residents. Internal consistency (Cronbach alpha = 0.73) and test-retest reliability (Lin correlation coefficient = 0.76) were good. There were significant differences between test scores and training levels ($P = .002$) and test scores correlated with in-service examination scores ($r = 0.38$; $P = .001$). There was no association between test scores and total number of cases or technical skills.

CONCLUSION: The Script Concordance Test appears to be a reliable, valid assessment tool for intraoperative decision-making during gynecologic surgery.

Key words: clinical competence, decision-making, education (medical, graduate), educational measurement, intraoperative judgment, operative/education/psychology, problem-based learning, surgical procedures

Cite this article as: 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;203:240.e1-6.

An obstetrics and gynecology resident's knowledge is assessed via multiple-choice examinations, while their critical thinking skills and judgment are assessed via oral examinations. However, there are no valid and reliable methods to assess a surgeon's judgment, ie, the ability to make decisions appropriate to an ambiguous clinical situation. For example, a resident may understand the indications for a surgical procedure (medical knowledge), and technically be able to perform the procedure (technical skill), but he or she may not be able to determine the appropriate next surgical step in an operative situation where several steps could be considered correct but one may be more appropriate to take before performing the others (surgical judgment).

Script Concordance Theory proposes a framework for assessing judgment by theorizing that clinicians initially process medical concepts via basic mechanisms of disease learned in medical school. As one progresses through training, a bank of cases based on engaging in patient care (ie, experience) is formed. One then creates knowledge and understanding to develop what have been referred to as "illness scripts."[1] According to this theory, people activate these previously acquired networks of knowledge when confronted with a new yet somewhat similar patient scenario. Expert clinicians develop this extensive pattern recognition based on experience working through the regular tasks and scenarios they encounter in practice.[2]

The Script Concordance Test (SCT) was developed using these principles to assess clinical reasoning skills for ill-defined clinical problems; the SCT has been used to assess judgment in surgery, urology, radiology, and internal medicine.[3-6] Compared to standard multiple-choice questions where there is 1 correct answer, scenarios using SCT are created where there is no single correct answer, and items about each scenario force trainees to confirm or eliminate clinical hypotheses based on their qualitative judgment in the face of an uncertain situation. Assuming that a surgeon's increasing experience improves judgment,

From the Department of Obstetrics and Gynecology (Drs Park, Barber, and Jelovsek), Cleveland Clinic, Cleveland, OH; the Department of Obstetrics and Gynecology (Dr Bent), IWK Health Center, Dalhousie University, Halifax, Nova Scotia, Canada; the Department of Obstetrics and Gynecology (Dr Dooley), Brooke Army Medical Center, San Antonio, TX; the Department of Obstetrics and Gynecology (Dr Dancz), Keck School of Medicine, University of Southern California, Los Angeles, CA; and the Department of Obstetrics and Gynecology (Dr Sutkin), University of Pittsburgh, Magee Women's Hospital, Pittsburgh, PA.

Presented orally at the 30th Annual Scientific Meeting of the American Urogynecologic Society, Hollywood, FL, Sept. 24-26, 2009.

Received Jan. 16, 2010; revised March 1, 2010; accepted April 8, 2010.

Reprints not available from the authors.

This research received a District V Research Grant from the American College of Obstetricians and Gynecologists and a Multi-Center Educational Grant and an Astellas Research Grant, both from the American Urogynecologic Society.

0002-9378/$36.00 • • doi: 10.1016/j.ajog.2010.04.010

a valid test should demonstrate that with increasing experience answers provided will come closer to those provided from a panel of experts. Such experience in a training program is usually demonstrated by increasing scores by resident training level.

The primary aim of this study was to develop a valid and reliable method to assess intraoperative judgment by obstetrics and gynecology resident trainees during gynecologic surgery using SCT. The resulting scores could then be used to assess the extent to which knowledge and understanding of gynecologic surgery meet the constraints and complexities of real-life complex operative scenarios that require the surgeon to apply his or her judgment. Secondary aims were to determine whether SCT scores correlate with other measures of knowledge and understanding such as standardized test scores from in-service examinations as well as technical skill as measured by the Global Rating Scale (GRS) of Operative Performance, surgical case numbers, and self-assessment of surgical judgment.

Materials and Methods

This was a multicenter, prospective study involving 5 obstetrics and gynecology residency programs in the United States and Canada: Cleveland Clinic (Cleveland, OH); Los Angeles County-University of Southern California Medical Center (Los Angeles, CA); IWK Health Center, Dalhousie University (Halifax, Nova Scotia, Canada); Brooke Army Medical Center (San Antonio, TX); and Magee Women's Hospital (Pittsburgh, PA). Institutional review board approval was obtained at all participating sites. All residents at all participating programs were approached for enrollment during their resident didactics by a study coordinator, informed consent was obtained, and trainees' identities were concealed using random number assignments.

The study was composed of 2 phases: phase 1 included item selection and test development and phase 2 included validity and reliability testing. Phase 1 involved developing the initial pool of scenarios for the SCT instrument, which were generated from the current literature on surgical education in gynecology and gynecologic oncology, as well as from the guidelines for resident education from the American College of Obstetricians and Gynecologists and the Council on Resident Education in Obstetrics and Gynecology (CREOG).[7] Three main principles to constructing a SCT were used: (1) an authentic challenging clinical situation is presented in which there are several relevant options; (2) responses follow 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 a series of experts.[8]

According to previous work by Charlin et al,[8] 50-60 items are sufficient to achieve internal reliability with an alpha coefficient of 0.80. Assuming scenarios or items would be eliminated, gynecologic surgeons from each participating site generated 96 case scenarios that occur during common gynecologic surgical procedures including 2-4 items per scenario. All scenarios were written to involve intraoperative decision-making and were designed to be ambiguous with no 1 correct answer. Common procedures were defined as those surgical procedures a resident should be able to perform determined by the CREOG Education Objectives.[7] The initial goal was to generate around 100 scenarios in anticipation of applying the following criteria to eliminate scenarios and items: >5% missing answers, minimum concordance >50%, maximum concordance <80%, items demonstrating floor or ceiling effects, or >80% agreement on experts' scoring on a particular answer.

Scenario design and scoring

The test design and scoring process is unique to SCT. Figure 1 is an example of a typical scenario with 3 items. Scenarios are constructed to reflect authentic challenging clinical situations, and multiple possible correct answers, which distinguish this questionnaire from the standard multiple-choice or oral examinations. The scoring process is based on the principle that any answer given by an expert has an intrinsic value, even if the other experts do not agree.[8] A target number of 10-20 experts was identified, based on a prior study demonstrating that at least 10 experts are necessary for acceptable reliability, and that >20 experts shows negligible additional benefit in terms of psychometric properties.[9] Credits for each answer are transformed proportionally to obtain a maximum score of 1 for answers most frequently endorsed by the expert panel for each item; other experts' choices are given partial credit. Answers not chosen by experts receive a score of 0. For example, if 6 of 10 experts choose response –1 to item no. 1 in Figure 1, this choice receives 1 point (6/6). If 3 experts choose response 0, this choice receives 0.5 point (3/6). If 1 expert chooses –2, this choice is assigned 0.167 (1/6). The responses +1 and +2 are assigned 0 points (Figure 2). The scoring system should reflect a range of potential scores and the expected distribution should be broad. However, ideally the distribution should also be clustered around a mean, because too broad a distribution invalidates the question.[6] Additionally, if only 1 answer is chosen by all the experts, the SCT becomes a multiple-choice question and should not be used. As such, this represented an additional item elimination criterion in the test development.

For this study, the scoring system was derived by answers generated by designated experts in gynecologic surgery from multiple sites (total n = 17; 2-6/site) using a weighted aggregate scoring method and Likert-type scale. Experts were selected by the principal investigator at each site for their operating experience and reputation. The total score for the test was the sum of credits on all items. For the convenience of interpretation, the scores were transformed so that the maximum score was 100.

Scores on the CREOG examination, a standardized, multiple-choice knowledge-based examination implemented yearly at most obstetrics and gynecology training programs in the United States and Canada, were collected to assess evidence of concurrent validity, or the relationship between knowledge and understanding and judgment. Surgical vol-

FIGURE 1
Example of single scenario with 4 items designed using Script Concordance Theory to assess intraoperative judgment in gynecologic surgery

A 55 year old female is undergoing an abdominal hysterectomy for menorrhagia. She has a 16 week size uterus with multiple fibroids on physical exam. During the hysterectomy you encounter bleeding from one uterine pedicle.

a) If you were considering treating the bleeding by cauterizing the bleeding vessel, and then you find the ureter immediately adjacent to the bleeding vessel, that treatment becomes (check one):

☐ -2	☐ -1	☐ 0	☐ +1	☐ +2
Almost contra-indicated or contra-indicated	Less appropriate or less useful	Neither more nor less appropriate	More appropriate or even helpful	Indicated or absolutely indicated

b) If you were considering treating the bleeding by transfusion, and then you find the hematocrit is 24 mg/dL, that treatment becomes (check one):

☐ -2	☐ -1	☐ 0	☐ +1	☐ +2
Almost contra-indicated or contra-indicated	Less appropriate or less useful	Neither more nor less appropriate	More appropriate or even helpful	Indicated or absolutely indicated

Park. Intraoperative judgment assessment using the SCT. Am J Obstet Gynecol 2010.

FIGURE 2
Scoring system for Script Concordance Test

	-2	-1	0	+1	+2
Number of experts choosing answer	1	6	3	0	0
Score	1/10	6/10	3/10	0	0
Transformed score	1/6	6/6	3/6	0	0
Credit per item	0.167	1	0.5	0	0

Park. Intraoperative judgment assessment using the SCT. Am J Obstet Gynecol 2010.

ume and performance of technical skill using a GRS of Operative Performance[10] were also used to investigate the relationship between performance of surgical skill and operative judgment. The GRS is a validated instrument to assess surgical technical skills in the operating room, and includes 7 domains covering respect for tissue, time and motion, instrument handling, knowledge of instruments, flow of operation, use of assistants, and knowledge of specific procedure.[10] The scale for each domain ranges from 1–5, with the maximum score of 35. Finally, trainees were asked to self-assess their own intraoperative judgment by marking an "X" on a 10-cm line on how one would rate his or her decision-making skills in the operating room as compared to an expert surgeon.

Internal consistency was evaluated by computing Cronbach alpha, and test-retest reliability was evaluated using Lin concordance correlation coefficient. Construct validity was evaluated by determining the questionnaire's ability to discriminate levels of training using Spearman correlation as presumably a trainee's judgment improves with increasing experience, particularly early in the learning experience. Test scores were correlated to CREOG scores, surgical volume, technical skills using GRS, and self-assessment of intraoperative judgment using Pearson correlation coefficient. A receiver operating characteristic (ROC) curve was generated to determine what threshold separated experts from nonexperts on performance of the test. Since we had a binary state of either resident or expert take the test and we had a continuous predictor variable on which to classify, the ROC curve was used. The threshold given denotes the score that most accurately predicts whether a resident reaches the level of judgment based on the test performance by the pool of experts. This score is not intended to be a minimum cutoff score to determine competence, as this would require different standard-setting methods.

Based on prior data looking at the validity of the SCT in assessing intraoperative decision-making skills in general surgery,[5] a sample size of 76 residents of a possible 124 residents from all participating sites was required to achieve 82% power to detect a 5% difference ± 6% (SD) between training levels using a 1-way analysis of variance with an alpha level of 0.05. All statistical analysis was performed on JMP 8.0 (SAS Institute, Cary, NC), SAS 9.1 (SAS Institute), and R 2.7.2 (R Foundation for Statistical Computing, Vienna, Austria) software.

RESULTS

Item reduction and test development (phase 1)

Item reduction was accomplished after administering 96 scenarios to 30 residents and 9 experts. Sixteen items were eliminated using the elimination criteria defined previously. The remaining 80 scenarios were reviewed for face and content validity by a separate panel of gynecologists, and 38 additional scenarios were deleted due to subject redundancy and/or failure to meet face and content validity. The final instrument included 42 scenarios with a total of 98 items.

TABLE 1
Number of residents at each participating site who took test and retest

Site	No. of residents who completed test 1	No. of residents who completed test 2
Cleveland Clinic, Cleveland, OH	25	11
University of Southern California, Los Angeles, CA	19	12
Magee Women's Hospital, Pittsburgh, PA	6	3
IWK Health Center, Dalhousie University, Halifax, Nova Scotia, Canada	13	9
Brooke Army Medical Center, San Antonio, TX	15	1
Total	78	36

Park. Intraoperative judgment assessment using the SCT. Am J Obstet Gynecol 2010.

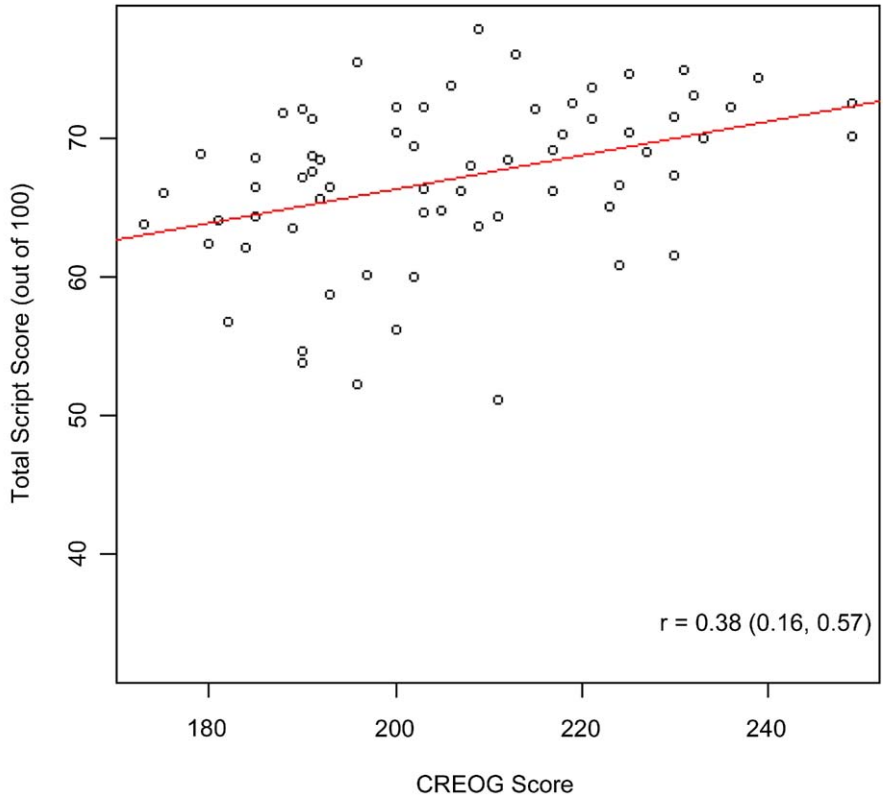
Reliability and validity testing (phase 2)

A final test, including 42 case scenarios for a total of 98 items, was administered to a total of 78 residents from the 5 participating sites: 14 (18%) postgraduate year (PGY)1, 19 (24%) PGY2, 20 (26%) PGY3, and 22 (28%) PGY4 residents. Three (4%) PGY5 residents from Canada took the questionnaire but were excluded from the analysis given the small number of residents at this training level. To assess test-retest reliability, the test was administered again to 36 residents 2-4 weeks after the initial administration. The number of residents who took the test and retest at each site is shown in Table 1. Seventeen experts also completed the final test including 5 generalist obstetrician-gynecologists, 3 gynecologic oncologists, 4 reproductive endocrinologists, and 5 urogynecologists. The median number of years in practice as 14.5 years (range, 2–24) and median cases per week was 4 (range, 2–15). There were no differences in expert scores by subspecialty ($P = .13$) or by site ($P = .96$), or any correlation of expert scores with years in practice ($P = .11$).

Mean total test time was 46.3 (±15.6) minutes for the residents and 50.8 (±25.7) minutes for the experts. The median resident score was 68.4 (mean, 66.8 ± 7.0) for the initial administration with a range of 32.5–77.8, and 67.8 (mean, 66.7 ± 7.1) for the retest (range, 51.1–75.9) with no significant differences in test scores between the 2 test episodes. GRS was available on 42 (56%) residents with a mean of 8 (±3.9) separate GRS collected per resident. The median GRS score was 29.0 (range, 21–35) of a total score of 35. Internal consistency was good (Cronbach alpha = 0.73) and test-retest reliability was 0.76 ($P < .001$) indicating strong agreement between the test and retest scores.

There was a significant difference in test scores by training level: mean examination scaled scores (0-100) by resident level were 63.7 (±6.8) for PGY1, 66.5 (±5.7) for PGY2, 65.9 (±5.8) for PGY3, and 68.6 (±5.8) for PGY4, and increasing scores correlated with increasing resident training level (r = 0.34; $P = .004$) demonstrating evidence in support of construct validity. Test scores also correlated with CREOG scores (r = 0.38; $P = .001$) (Figure 3). There were no significant correlations between resident test scores and total number of surgical cases, skill performance based on the GRS scores, or self-assessment of judgment (Table 2).

FIGURE 3
Script Concordance Test score correlation with Council on Resident Education in Obstetrics and Gynecology (CREOG) score



$P = .001$

Park. Intraoperative judgment assessment using the SCT. Am J Obstet Gynecol 2010.

A ROC curve was generated to determine a cutoff score in which a trainee has achieved scores similar to the large majority of experts (Figure 4). A total score of 73 was associated with an area under the curve of 0.87 (95% confidence interval, 0.78–0.97) indicating that when a trainee scores ≥73 on the test they have reached similar performance of experts taking the same test. Twelve of 17 (70.6%) experts scored above this threshold, while 10 of 75 (13.3%) residents scored above the threshold including 1 of 13 (7.7%) first-year residents, 1 of 18 (5.6%) second-year residents, 1 of 19 (5.3%) third-year residents: and 6 of 20 (30%) fourth-year residents.

TABLE 2
Correlations of resident test scores with experience indicators

Factor 1	n[a]	Correlation	95% CI	*P* value
CREOG score	67	0.38	0.16–0.57	.001
GRS	39	0.19	–0.14 to 0.48	.25
Total cases	70	0.15	–0.09 to 0.37	.21
Self-assessment of judgment	68	0.12	–0.12 to 0.35	.31
Years of training	70	0.34	0.11–0.56	.004

CI, confidence interval; *CREOG*, Council on Resident Education in Obstetrics and Gynecology in-service examinations; *GRS*, Global Rating Scale of Operative Performance.

[a] No. of residents with available relevant information. Some residents did not provide some data, eg, GRS, total no. of cases, self-assessment of judgment, or years in training (ie, postgraduate year status). Not all residents had CREOG scores available.

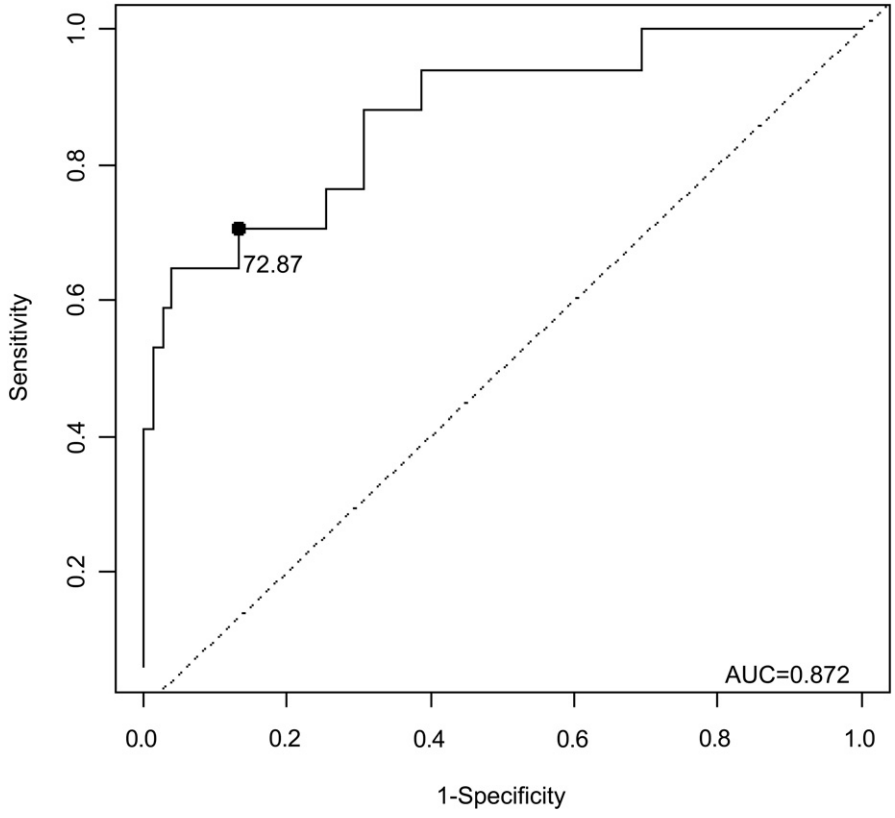
Park. Intraoperative judgment assessment using the SCT. Am J Obstet Gynecol 2010.

COMMENT

This study demonstrates that Script Concordance Theory may be used to assess an obstetrics and gynecology resident's judgment in the operating room. This test appears to be internally consistent, have good test-retest reliability, and have evidence of construct validity supported by increasing scores with increasing training experience and CREOG performance. Our findings are consistent with previous studies in other specialties demonstrating construct validity of SCT with scores increasing with higher level of training.[4-6] For example, a cohort of 24 students from a medical school in Quebec, Canada, was followed up until the end of their family medicine residency training; the SCT highly correlated with short answer management problems and simulated office oral examinations.[11] Meterissian et al[4,5] demonstrated that the SCT can discriminate between junior and senior general surgery residents while other studies have investigated similar tests for residents in surgery, urology, and radiology.[3,6] The SCT design also appears to be stable across 2 different linguistic and learning environments as demonstrated by French and Canadian urology residents.[12,13] Combined, these results corroborate the validity of the aggregate scoring methods used by SCT.

One of this study's strengths is the multicenter, North American sample. Additionally, these SCT scores were compared to standardized test scores of knowledge and understanding, operative experience, operative skill, and self-assessment of judgment. It was not surprising that a surgeon's judgment correlated to his or her specialty-specific knowledge and understanding given that both are a required component of forming proper judgment. Likewise, it was not surprising that operative skill did not correlate with increasing judgment, although a potential reason for not finding a correlation is the lack of precision in assess-

FIGURE 4
Receiver operating characteristic (ROC) curve demonstrating cutoff score of 73 at which trainee reaches similar performance as experts taking test



AUC, area under curve.

Park. Intraoperative judgment assessment using the SCT. Am J Obstet Gynecol 2010.

ing surgical skills, decreasing our ability to find an association that may truly exist. Knowing what to do, even in complex intraoperative scenarios, does not necessarily equate with a surgeon's ability to perform a skill under those circumstances. Similarly, understanding or mastering the technical aspects of a procedure does not necessarily imply good surgical judgment. Our findings provide support for assessing knowledge, skill, and judgment independently since how well one is able to perform a skill is separate from the decisions to perform it or other complex intraoperative decisions that might be required. Finally, this study provides a practical cutoff score that a trainee or program director could present as a useful target for trainees to reach during their training or that could be used in maintenance of certification to continuously assess a practicing surgeon's judgment with the changing values in the medical environment. However, this cutoff score was not determined using formal standard-setting criteria typically used in educational testing and therefore really should only be used as a screening tool to identify candidates for remediation not for high stakes decision-making, such as whether or not a resident should advance to the next level or as a criteria for graduation. Limitations include the test's narrow focus on intraoperative judgment. Clinical decision-making regarding whether to operate at all (preoperative) or judgment managing postoperative scenarios, or office-based or obstetric cases were not assessed in this test. Moreover, the study sample only included large training programs making the test characteristics unknown in smaller community training programs.

CREOG examinations use a multiple-choice format and provide a reliable and valid assessment of a resident's knowledge and ability to apply that knowledge to a relevant clinical scenario. However, a limitation to multiple-choice formats is the cueing effect in which the right answer is always presented to the examinee. This is not the case in real life and does not apply using the SCT format. Different question formats are useful to assess outcomes where some questions have 1 correct answer and some questions have multiple correct answers. Since there is no 1 superior question format, it is generally recommended that combinations of formats are used to minimize the cueing effect and as long as these approaches are feasible. For example, residency program directors could use CREOG examination scores to assess lower-order cognition such as knowing facts and application of that knowledge to standardized scenarios, and use an examination in the SCT format for higher-order cognition to examine application of knowledge in complex real-life scenarios that do not have a single correct answer. Directors could also use both assessments to provide remediation to residents who are falling behind. Although the SCT format appears to contain important preliminary psychometric properties to be used as a method of assessing judgment and is much less personnel intensive to administer than in-person oral examinations, its role as a supplement to or replacement of oral examinations for high stakes testing or maintenance of certification remains to be determined.

Given the restrictions on resident work hours, declining surgical volume, and increasing scrutiny by professional societies and the public, it is imperative that training programs verify that their trainees achieve competent performance. A valid and reliable method of assessing judgment such as the SCT may be useful since it offers a standardized method to assess judgment. In the future, these scenarios could be incorporated into simulation models and simulated scenarios, which are increasingly being adopted as part of the accreditation process for many training programs. In summary, scores on an intraoperative assessment of judgment during gynecologic surgery are higher for individuals who have more experience. This SCT appears to be a reliable, valid instrument for assessing intraoperative decision-making during gynecologic surgery. ■

ACKNOWLEDGMENT

We thank Begum Ozel, MD.

REFERENCES

1. Schmidt HG, Norman GR, Boshuizen HP. A cognitive perspective on medical expertise: theory and implication. Acad Med 1990;65:611-21.

2. 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.

3. Sibert L, Darmoni SJ, Dahamna B, Hellot MF, Weber J, Charlin B. On line clinical reasoning assessment with script concordance test in urology: results of a French pilot study. BMC Med Educ 2006;6:45.

4. Meterissian S, Zabolotny B, Gagnon R, Charlin B. Is the script concordance test a valid instrument for assessment of intraoperative decision-making skills? Am J Surg 2007;193:248-51.

5. Meterissian SH. A novel method of assessing clinical reasoning in surgical residents. Surg Innov 2006;13:115-9.

6. Charlin B, Brailovsky CA, Brazeau-Lamontagne L, Samson L, Van der Vleuten CP. Script questionnaires: their use for assessment of diagnostic knowledge in radiology. Med Teach 1998;20:567-71.

7. Council on Resident Education in Obstetrics and Gynecology. Educational objectives: a core curriculum in obstetrics and gynecology, 8th ed. Washington, DC: American College of Obstetricians and Gynecologists; 2005.

8. Charlin B, Roy L, Brailovsky C, Goulet F, van der Vleuten C. The script concordance test: a tool to assess the reflective clinician. Teach Learn Med 2000;12:189-95.

9. Gagnon R, Charlin B, Coletti M, Sauve E, van der Vleuten C. 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.

10. Reznick R, Regehr G, MacRae H, Martin J, McCulloch W. Testing technical skill via a innovative "bench station" station. Am J Surg 1997;173:226-30.

11. Brailovsky C, Charlin B, Beausoleil S, Cote 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 experimental study on the script concordance test. Med Educ 2001;35:430-6.

12. Sibert L, Charlin B, Corcos J, Gagnon R, Lechevallier J, Grise P. 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.

13. Sibert L, Charlin B, Corcos J, Gagnon R, Grise P, van der Vleuten C. Stability of clinical reasoning assessment results with the script concordance test across two different linguistic, cultural and learning environments. Med Teach 2002;24:522-7.