

Assessing intraoperative judgment using script concordance testing through the gynecology continuum of practice*

NATHAN KOW¹, MARK D. WALTERS¹, MICKEY M. KARRAM², CARLOS J. SARSOTTI³ & J. ERIC JELOVSEK¹

¹Cleveland Clinic, USA, ²Christ Hospital, USA, ³Italian Hospital of Buenos Aires, Argentina

Abstract

Objective: To measure surgical judgment across the Obstetrics and Gynecology (OBGYN) continuum of practice and identify factors that correlate with improved surgical judgment.

Methods: A 45-item written examination was developed using script concordance theory, which compares an examinee's responses to a series of "ill-defined" surgical scenarios to a reference panel of experts. The examination was administered to OBGYN residents, Female Pelvic Medicine and Reconstructive Surgery (FPMRS) fellows, practicing OBGYN physicians and FPMRS experts. Surgical judgment was evaluated by comparing scores against the experts. Factors related to surgical experience were measured for association with scores.

Results: In total, 147 participants including 11 residents, 37 fellows, 88 practicing physicians and 11 experts completed the 45-item examination. Mean scores for practicing physicians (65.2 ± 7.4) were similar to residents (67.2 ± 7.1), and worse than fellows (72.6 ± 4.2 , $p < 0.001$) and experts (80 ± 5 , $p < 0.001$). Positive correlations between scores and surgical experience included: annual number of vaginal hysterectomies ($r = 0.32$, $p = < 0.001$), robotic hysterectomies ($r = 0.17$, $p = 0.048$), stress incontinence ($r = 0.29$, $p < 0.001$) and prolapse procedures ($r = 0.37$, $p < 0.001$). Inverse correlation was seen between test scores and years in practice. ($r = -0.19$, $p = 0.02$).

Conclusion: Intraoperative judgment in practicing OBGYN physicians appears similar to resident physicians. Practicing physicians who perform FPMRS procedures perform poorly on this examination of surgical judgment; lower performance correlates with less surgical experience and the greater amount of time in practice.

Introduction

Our current system of training surgeons relies heavily on experiences during postgraduate training. With continuing changes in work hours, advances in surgical technology, and emphasis on quality of care and eliminating errors, verification of a graduating resident's competence performing surgery is undergoing increased scrutiny. The goal of training programs is to equip surgeons with specialty-specific knowledge and necessary diagnostic and therapeutic procedural skills through a process of direct supervision, repetitive practice, and assessment by faculty. While learning the technical aspects of surgery is important, it is estimated that approximately 75% of significant events are related to judgment (Spencer 1978). It is assumed, even if it is rarely measured, that through the postgraduate experience medical and surgical judgment improves with a trainee ultimately acquiring enough judgment to be a safe, independent practitioner. In the United States, undergraduate medical graduates students enter into

Practice points

- Surgical judgment varies throughout the continuum of gynecologic practice;
- While experience is important in the attainment and maintenance of judgment, surgical experience correlates with higher judgment, while amount of time in practice correlates with lesser judgment.
- Future studies using assessment strategies to understand, maintain and improve surgical judgment should be performed to ensure safe, effective surgical care.

their residency for general training in a chosen specialty after completing general licensing examinations. They may choose to further their training through a subspecialty fellowship. Prior to matriculation from residency, many specialties require their trainees to sit for an examination consisting of a written test that reliably tests a physician's ability to apply medical

Correspondence: Nathan Kow, MD, Fellow, Female Pelvic Medicine and Reconstructive Surgery, Division of Female Pelvic Medicine & Reconstructive Surgery, A81, Obstetrics, Gynecology, & Women's Health Institute, Cleveland Clinic, 9500 Euclid Avenue, Cleveland, OH 44195, USA. Tel: 216-445-6586; Fax: 216-636-2620; E-mail: kown@ccf.org

*This article was presented at the 2013 CREOG & APGO Annual Meeting, February 27–March 2, 2013, Phoenix, Arizona and also at 2013 International Continence Society Meeting, August 26–30, 2013, Barcelona, Spain.

knowledge. After some time in practice, many examinees take an oral examination primarily designed to assess clinical decision making and judgment. Certification in Obstetrics and Gynecology (OBGYN) in the United States requires oral board examinations which represent the final formalized assessment of judgment while additional processes such as maintenance of certification and peer-review methods are primarily designed to maintain a surgeon's knowledge base and promote safe care. Meanwhile, it is assumed that a surgeon's judgment improves but is not currently subjected to formal evaluation.

Our current understanding of the natural course of surgical judgment after surgeons complete their training program is not well understood. While logic and some anecdotal evidence suggests that continued "practice of surgery" through surgeons' cases help to refine the network of experiences for which surgeons can apply to newly encountered scenarios, a recent meta-analysis of published data suggests that this is not necessarily the case. In fact, published data suggest that the quality of healthcare provided may actually decrease over time across both medical and surgical practices (Choudhry et al. 2005). These findings highlight the need to develop tools to assess and study the natural course of judgment in practicing surgeons. Understanding judgment through a surgeon's career of practice would allow for the design of appropriate educational interventions and processes and should result in improved quality of healthcare.

An emerging alternative to assess clinical decision making or judgment uses a novel multiple-choice written examination based on Script Concordance Theory. First developed by Charlin et al., Script Concordance Testing (SCT) aims to compare an examinee's responses to a series of "ill-defined" scenarios in which correct decisions are weighted and compared to a reference panel of experts. These scenarios can be diagnostic, prognostic or therapeutic by design. SCTs have been previously validated in Surgery, Urology, Internal Medicine and OBGYN (Sibert et al. 2002; Brazeau-Lamontagne et al. 2004; Meterissian 2006; Meterissian et al. 2007; Park et al. 2010).

Our primary aim was to measure surgical judgment using a written examination using Script Concordance Theory across the OBGYN continuum of practice. Our hypothesis was that surgical judgment depends more on surgical volume than years in practice and that this would be reflected between groups of surgeons at different points along the continuum of practice. Secondary aims were to identify factors that correlated with intraoperative judgment and to assess reliability of the examination in this study population.

Materials and methods

This was a cross-sectional study involving OBGYN residents, Female Pelvic Medicine and Reconstructive Surgery (FPMRS) fellows, practicing OBGYN physicians and FPMRS experts. Subspecialty fellows were defined as participants who completed a residency training program in OBGYN and were currently enrolled in an additional training program

in FPMRS. Practicing physicians were defined as any participant who had successfully completed a residency training program and/or fellowship training program practicing without supervision. Subspecialty fellow and practicing surgeon attendees enrolled in continuing medical education courses offered through the International Academy of Pelvic Surgery between February 2012 and August 2012. Courses were held in both the United States and Argentina with a focus on pelvic organ prolapse and urinary incontinence. Resident participants were defined as trainees in general OBGYN and were recruited from the Cleveland Clinic OBGYN Residency Program. Experts were recognized academic surgeons in the field of FPMRS and were additionally course instructors at the CME courses. Ethical approval of the study was obtained through the Cleveland Clinic Institution Review Board and informed consent was not required from participants.

A test specific for assessing intraoperative judgment in the field of FPMRS was designed. The format of this examination was designed from similar cases in general OBGYN which have been demonstrated to be reliable and valid in a multi-center setting (Park et al. 2010). The domains of the test covered intraoperative management of surgical complications or abnormal findings at the time of female pelvic reconstructive surgery. The test consisted of case stems that provide a surgical vignette that, if real, would require further surgical therapeutic action. Each vignette was followed by new additional information and is followed by a question that asks the participant, given the new information, whether they are more or less likely to proceed with their original surgical strategy (Figure 1). A total of 20 case stems, each with 2–3 response items were constructed for a total of 45 items. Face validity of the items were obtained by allowing additional experts to view and give feedback regarding case stems and items before administration of the final test. The length of the test was also limited to 20 cases as it was felt this was the upper limit of questions that could be asked in a CME setting to practicing surgeons without affecting response rates (Park et al. 2010).

All course attendees were approached for enrollment at the beginning of the course. Attendees agreeing to participate were given a study identifier in order to conceal their identity. Participant characteristics and practice information including years in practice; attainment of fellowship training; age; type of practice (academic or private); annual number of vaginal, abdominal, laparoscopic and robotic hysterectomies; annual number of pelvic organ prolapse procedures and incontinence procedures; and urodynamic (UDS) and cystoscopy privileges were collected on all subjects. The Script Concordance Test – Female Pelvic Medicine and Reconstructive Surgery (SCT-FPMRS) was originally created in English. Translation of the English version into a Spanish version was performed by professional translators (H-I Translating & Interpreting, LLC). Translated tests were independently reviewed by surgeons in their native language prior to administration to subjects for any clarifications. The SCT-FPMRS test was administered to participants at the beginning of the conference in order to obtain a baseline score and practicing physicians were asked to retake the test two days later in order to test for test-retest reliability.

You are performing a retropubic midurethral sling on a 65 year old women with stress urinary incontinence. After placement of the trocars, you perform a cystoscopy and notice a perforation of the trocar through the bladder wall.					
	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
If you remove the trocar and notice bleeding from the puncture site, sling placement becomes:	-2	-1	0	+1	+2
If you also notice a significant amount of bleeding from the suprapubic skin puncture sites, sling placement becomes:	-2	-1	0	+1	+2

Figure 1. Example of clinical vignette.

Scoring and establishment of expert panel

Scoring was calculated for each case using a modified aggregate method. Experts individually completed the test and their answers were used to build the scoring key. This is a unique feature of SCT as scenarios are specifically designed not to have a single best answer, but rather to have a range of correct answers. This property acknowledges an important reality in surgical practice: that even experienced surgeons interpret data, make judgments, and respond to uncertain surgical situations in a variety of ways (Lubarsky et al. 2013). Individual expert scores were calculated based on a scoring key that excludes their own responses to avoid bias. Partial credit was assigned to each response based on how many of the expert panel members chose that response. Each question received a credit of 1 point if the answer matched the modal answer from the expert panel. A score of 0 was assigned if the participant's answer did not match any of the expert panel's answers. Individual question scores were then averaged within a particular case in order to calculate a case score and the case scores were totaled to obtain the total test score. The max score for the test equals the total number of cases; in this study the total cases equals 20. There were no deviate responses in the expert panel. Since all expert scores were within 2 standard deviations (SD) of the mean, all were included as part of the final panel (Charlin et al. 2010).

It is common to standardize test scores by providing the deviation of a participant's test scores from the mean of all test takers. Since SCT testing involves reporting a final score that reflects the difference in performance between the participant and the expert panel, it is necessary to transform the expert panel scores to provide a reference. Participant's tests were then scored based on the case scores from the expert panel. Briefly, z scores were calculated using the mean and SDs of the established expert panel set at 0 and 1. Subsequently, z scores were transformed based on a panel mean and SD

of 80 and 5, providing the final t scores to be compared. Transform scores were utilized in order to help with interpretation of test scores. Raw scores were calculated for the participants based on the sum of credits on all cases (maximum 20). Raw scores were transformed based on the mean and SD of the expert panel.

Participants were included for analysis if they completed greater than 80% of the items. Missing answers were multiply imputed to ensure complete data. Ten imputations were performed with the mean of the imputed responses being accepted as the response value. Comparisons of test scores against the expert panel scores were performed using student t -tests. Multiple between group comparisons were adjusted using the Bonferonni correction. Additional measures of surgical experience including years of practice and annual number of procedures were performed using Pearson's correlation coefficient. Internal consistency and test-retest reliability were evaluated using participants who provided complete data. Internal consistency was evaluated by computing Cronbach α . Test-retest reliability is measured with the Intraclass Correlation Coefficient.

Results

A total of 147 participants, including 11 residents, 37 fellows, 88 practicing physicians and 11 experts completed the 45-item examination. Participant characteristics and surgical experience are summarized in Table 1. Median (range) age for residents, fellows, practicing physicians and experts were 29 (ranges), 33, 41, and 46, respectively. The majority of resident and fellows were female reflecting the current distribution in U.S. OBGYN training programs. Male participants represented the majority of practicing physicians and experts. Practicing physicians primarily worked in the private practice setting compared with residents, fellows, and experts who worked in the academic setting. In general, annual

Table 1. Participant characteristics and indicators of surgical experience.

	Resident	Fellow	Physician	Expert
Age ^a	29 [27,30]	33 [31,35]	42 [36,50]	46 [41,51]
Sex ^b				
Male	1 (9)	8 (22)	59 (67)	6 (55)
Female	10 (91)	29 (78)	29 (33)	5 (45)
Years in practice ^a	2 [1,3]	6 [6,6]	12 [6,20]	12 [9-5,19]
Type of practice ^b				
Private practice	0 (0)	2 (5-7)	51 (59-3)	1 (9)
Academic	11 (100)	33 (94-3)	12 (14)	10 (91)
Both	0 (0)	0 (0)	23 (26-7)	0 (0)
Annual number of surgeries				
Vaginal hysterectomy ^c	2.8 (5-7)	32.8 (24-8)	19.9 (24-7)	73.2 (6-2)
Abdominal hysterectomy ^c	3.3 (7-1)	5.9 (7-4)	14.9 (18-2)	5 (6-2)
Laparoscopic Hysterectomy ^c	2.2 (6-7)	15.1 (23)	10.7 (20-5)	12.9 (14-4)
Robotic hysterectomy ^c	0.2 (0-7)	13.7 (18-6)	4.2 (12-6)	6.7 (11-8)
Stress incontinence ^c	1.1 (3-3)	78.3 (36-3)	39 (38-3)	107.7 (43-5)
Pelvic organ prolapse ^c	1.1 (3-3)	86.7 (45-7)	32.9 (31-8)	125.4 (59-9)
UDS privileges ^b				
No	n/a	n/a	52-87	0
Yes	n/a	n/a	47-13	100
Cystoscopy privileges ^b				
No	n/a	n/a	13-79	0
Yes	n/a	n/a	86-21	100

^aMedian [P25,P75]; ^bN (%); ^cMean (SD).

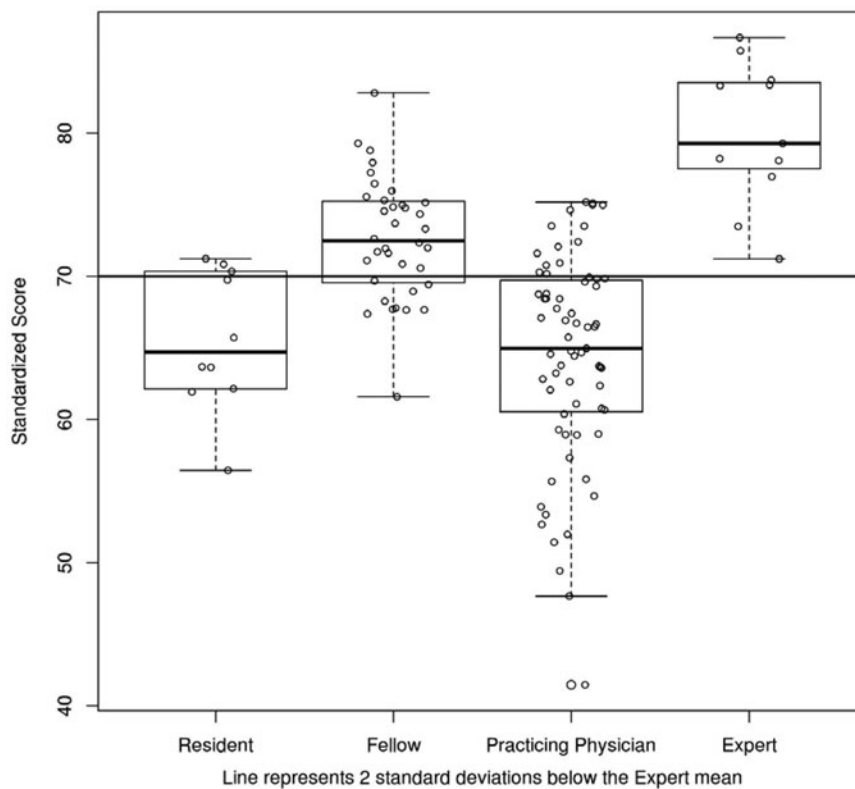


Figure 2. Scores summary.

surgical case load was higher in fellows and experts compared with residents and practicing physicians.

Mean transformed scores for practicing physicians (65.2 ± 7.4) were similar to residents (67.2 ± 7.1), and significantly less than fellows (72.6 ± 4.2 , $p < 0.001$) and experts (80 ± 5 , $p < 0.001$) (Figure 2). For SCT, it has been proposed that the pass mark should be determined in relation to the performance of experts who have sat the same test

and that performance ≥ 2 SDs from the experts would be considered a failure of the test (Lubarsky et al. 2013). The distribution of scores of residents, fellows and practicing physicians compared to expert scores are shown in Table 2. The majority (82%) of practicing physicians scored ≥ 2 SDs below the mean expert score. Mean differences in scores were compared between groups using Bonferonni correction showed that all groups were statistically different except

Table 2. Distribution of test scores relative to expert scores.

	N	Mean (SD)	Below 2 SD N (%)	Below 1 SD N (%)	From 1 SD to mean N (%)	Above the mean N (%)
Expert	11	80.0(5.0)				
Resident	11	67.2(7.1)	7 (63.6)	3 (27.3)	0 (0)	1 (9.1)
Fellow	37	72.6(4.3)	11 (29.7)	16 (43.2)	9 (24.3)	1 (2.7)
Practicing Physician	88	65.2(7.4)	72 (81.8)	12 (13.6)	2 (2.3)	2 (2.3)

for residents and practicing physicians (mean diff=1.74, $p=0.42$). The difference between fellow and resident scores also did not reach statistical significance, but trended towards the fellow (mean diff=5.40, $p=0.02$). When groups were compared with practicing physicians, the mean difference in scores was similar to residents (1.74, $p=0.42$) and significant higher in fellows (7.14, $p<0.001$) and experts (14.6, $p<0.001$).

There were significant correlations between test scores and additional measures of surgical experience including number of annual vaginal hysterectomies ($r=0.32$, $p<0.001$), robotic hysterectomies ($r=0.17$, $p=0.048$), stress incontinence procedures (0.29, $p<0.001$) and pelvic organ prolapse procedures (0.37, $p<0.001$). There were also significant negative correlations between test scores and years in practice ($r=-0.19$, $p=0.02$) and annual number of abdominal hysterectomies ($r=-0.23$, $p=0.006$). Overall, examination internal consistency (Cronbach's $\alpha=0.72$) and test-retest reliability (ICC=0.67) were acceptable.

Discussion

This study demonstrates a concerning finding of low surgical judgment in practicing OBGYN physicians. It also demonstrates that surgical volume, not years in practice, correlates with better intraoperative judgment. In fact, intraoperative judgment using this tool appeared to negatively correlate with years in practice and perform at a level similar to residents. Based on these findings it may be that annual surgical volume of appropriate surgeries should be considered as one of the surrogates of surgical judgment rather than years of surgical practice. In the absence of a formal examination of judgment, surgical volume could be used as this surrogate to determine the level of judgment. This supports script concordance theory in which it is believed that clinical or surgical decision-making and judgment are improved by developing and continuing to refine a bank of well defined "scripts" that can be drawn upon when a surgeon encounters an "ill-defined problem (Charlin et al. 2000)." A progressive improvement of scores from resident to fellows to practicing surgeons supports the evidence that judgment improves through training, peaks shortly after training, and then begins to plateau or decline. This has been shown in several studies by demonstrating an increase in mortality of patients undergoing cardiac procedures with increasing years of practice of the primary surgeon (Hartz et al. 1999; O'Neill et al. 2000). Similar findings of increased mortality is also seen in a study of outcomes of acute myocardial infarction that showed an increase of

0.5% mortality for every year since the physician graduated from medical school (Norcini et al. 2000).

These findings also suggest that surgeons should continue to operate in order to maintain or improve judgment as test scores correlate with greater surgical volume. Additionally, this study supports the notion that judgment needs to be reevaluated in the setting of a low surgical volume practice. Hospital privileging may need to incorporate the recommendation that surgeons should maintain a minimum annual case load particularly for more complex or specialized treatments. Similarly, recent recommendations by the OBGYN Residency Review Committee in the U.S. are beginning to ensure that trainees achieve a minimum number of surgical procedures to attain adequacy for independent practice. These data suggest that such a minimum number may need to be calculated for practicing physicians. The addition of a reliable metric for surgical judgment would provide further evidence of competence.

Since it is likely that the changing landscape of training programs will continue to present barriers to achieving competence in surgical procedures, additional factors such as work-life balance and gender changes in the work force may also impact surgical volume. In a univariable analysis of gender, women scored slightly higher than men (Mean score 70 versus 66.9, $p=0.01$). However, on a multivariable model, this difference was dampened due to the larger effect of the other variables making gender less important with respect to their SCT score ($p=0.44$).

Strength of this study is the use of a large group of practicing physicians who provide a wide range of surgical experience. SCT development and validation has occurred primarily in training programs because there are clear demarcations between each post-graduate level that can be utilized to obtain construct validity. This has resulted in multiple previous studies showing improvement in clinical judgment as trainees progress through their residency program. However, little is known regarding how judgment changes with practicing physicians and this study adds to this gap. This study is one of the largest studies to investigate surgical judgment through the entire continuum of practice.

One of the main limitations in this study is that surgical volume could be a confounder. We do not know whether maintaining high surgical volume results in a maintenance or increase in surgical judgment or whether surgeon's with good judgment tend to get more patients referred to them for surgery. We suspect that it is a combination of both factors. Furthermore, as the test focused on case scenarios seen in FPMRS, it is unclear what effect this has on participants who only have general training in OBGYN but practice in what is an emerging subspecialized area in OBGYN in the United

States. Attendance of the courses may indicate their desire to improve upon a known deficiency in their practice and could explain the lower scores seen in the practicing physician group. Conversely, the majority of participants indicated that they did perform pelvic reconstructive procedures based on their self-reported annual surgical volume. Therefore, it would be expected that they have knowledge of the procedures and potential complications giving them a potential advantage to generalists who do not perform these procedures. A potential explanation is that while knowledge of the procedures is important, additional factors are required to improve judgment. Alternatively, this finding could represent the natural course of judgment. This highlights the need for the design of programs to assist in maintenance and improvement in judgment.

While SCT has been validated in multiple medical and surgical specialties, it not yet been widely implemented as an evaluation tool in training programs or in formalized evaluation of practicing physicians. Given the current training and practice environment, the addition of measures of judgment would enhance our ability to evaluate trainees and practicing physicians. Current testing modalities utilizing oral examinations and case list preparation is both time consuming for examinees and for examination centers. The man power required to be able to administer oral examinations is significantly greater than a written examination. Therefore, SCT may have the potential to supplement or replace our current evaluation system and allow for interim evaluation of practicing physicians. Future studies using assessment strategies to understand, maintain and improve surgical judgment should be performed in order to ensure safe and effective surgical care.

Notes on contributors

NATHAN KOW, MD, is a Fellow in Female Pelvic Medicine & Reconstructive Surgery at the Center of Urogynecology and Reconstructive Pelvic Surgery, Cleveland Clinic, Cleveland, Ohio.

MARK D. WALTERS, MD, is a Professor and Vice Chair of the Center of Urogynecology and Reconstructive Pelvic Surgery and the Obstetrics, Gynecology & Women's Health Institute, Cleveland Clinic, Cleveland, Ohio.

MICKEY M. KARRAM, MD, is a Professor in the Division of Female Pelvic Medicine & Reconstructive Surgery, Christ Hospital, Cincinnati, Ohio.

CARLOS J. SARSOTTI, MD, is a Professor, Urogynecology Section, Italian Hospital of Buenos Aires, Buenos Aires, Argentina.

J. ERIC JELOVSEK, MD, is an Associate Professor and Residency Director at the Center of Urogynecology and Reconstructive Pelvic Surgery, and also serves as the residency director of the Obstetrics, Gynecology & Women's Health Institute, Cleveland Clinic, Cleveland, Ohio.

Acknowledgements

We would like to acknowledge Cheryl Perrero for all her work assisting us in coordinating the administration of the tests

during the CME courses provided by the International Academy of Pelvic Surgery. We would also like to acknowledge Henry Castro, MD for his assistance in proofing the Spanish version of the test.

Declaration of interest: Nathan Kow, Mark D. Walters, and J. Eric Jelovsek have no disclosures. Mickey Karram serves as a speaker and consultant for Astellas, Medtronic, and American Medical Systems. Additionally, he receives royalties from Emedsco. Carlos Sarsotti is a proctor for the Latin American branch of American Medical Systems. The authors alone are responsible for the content and writing of the article.

References

- Brazeau-Lamontagne L, Charlin B, Gagnon R, Samson L, van der Vleuten C. 2004. Measurement of perception and interpretation skills during radiology training: Utility of the script concordance approach. *Med Teach* 26(4):326–332.
- Charlin B, Gagnon R, Lubarsky S, Lambert C, Meterissian S, Chalk C, Goudreau J, van der Vleuten C. 2010. Assessment in the context of uncertainty using the script concordance test: More meaning for scores. *Teach Learn Med* 22(3):180–186.
- Charlin B, Tardif J, Boshuizen HP. 2000. Scripts and medical diagnostic knowledge: Theory and applications for clinical reasoning instruction and research. *Acad Med: J Assoc Am Med Coll* 75(2):182–190.
- Choudhry NK, Fletcher RH, Soumerai SB. 2005. Systematic review: The relationship between clinical experience and quality of health care. *Ann Intern Med* 142(4):260–273.
- Hartz AJ, Kuhn EM, Pulido J. 1999. Prestige of training programs and experience of bypass surgeons as factors in adjusted patient mortality rates. *Med Care* 37(1):93–103.
- Lubarsky S, Dory V, Duggan P, Gagnon R, Charlin B. 2013. Script concordance testing: From theory to practice: AMEE Guide No. 75. *Med Teach* 35(3):184–93.
- Meterissian S, Zabolotny B, Gagnon R, Charlin B. 2007. Is the script concordance test a valid instrument for assessment of intraoperative decision-making skills? *Am J Surg* 193(2):248–251.
- Meterissian SH. 2006. A novel method of assessing clinical reasoning in surgical residents. *Surg Innov* 13(2):115–119.
- Norcini JJ, Kimball HR, Lipner RS. 2000. Certification and specialization: Do they matter in the outcome of acute myocardial infarction? *Acad Med: J Assoc Am Med Coll* 75(12):1193–1198.
- O'Neill L, Lanska DJ, Hartz A. 2000. Surgeon characteristics associated with mortality and morbidity following carotid endarterectomy. *Neurology* 55(6):773–781.
- Park AJ, Barber MD, Bent AE, Dooley YT, Dancz C, Sutkin G, Jelovsek JE. 2010. Assessment of intraoperative judgment during gynecologic surgery using the Script Concordance Test. *Am J Obstet Gynecol* 203(3):240.e1–240.e6.
- Sibert L, Charlin B, Corcos J, Gagnon R, Lechevallier J, Grise P. 2002. Assessment of clinical reasoning competence in urology with the script concordance test: An exploratory study across two sites from different countries. *Eur Urol* 41(3):227–233.
- Spencer F. 1978. Teaching and measuring surgical techniques: The technical evaluation of competence. *Bull Am Coll Surg* 63:9–12.