Research article Open Access

Bi-cultural, bi-national benchmarking and assessment of clinical reasoning in Obstetrics and Gynaecology

Paul Duggan[2], Patricia Monnier[3], Alphonse Roex[2], Marie-Josée Bédard[4], Bernard Charlin[5]

Corresponding author: Prof Paul Duggan paul.duggan@adelaide.edu.au

Institution: 2. Discipline of Obstetrics and Gynaecology, University of Adelaide, 3. The Research Institute of the McGill University Health Centre, 4. Le Centre hospitalier de l'Université de Montréal , 5. Centre for Pedagogy Applied to the Health Sciences (CPASS), University of Montreal

Categories: Assessment, Students/Trainees, Teaching and Learning

Received: 20/06/2016
Published: 23/06/2016

Abstract

Background: The Script Concordance Test (SCT) is being increasingly used in professional development in clinical reasoning (CR) in postgraduate medicine. On-line delivery favours multi-institutional collaboration.

Objectives: To establish if: 1) SCT questions developed in the French-speaking University of Montreal were readily adaptable for use in the English-speaking University of Adelaide 2) expert reference panels (ERP) from both institutions could be used interchangeably 3) student cohorts would perform similarly in the same test.

Study Design: 82 SCT questions based on 27 clinical cases in Obstetrics and Gynaecology were developed in Montreal and run in a volunteer cohort of year 3 and year 4 medical students (n=154). Local faculty translated all questions, selecting 31 based on 17 clinical cases for use in summative examinations a year 5 student cohort in Adelaide (n=123).

Results: Mean (SD) percentage scores using each ERP key were: 74.2 (6.4) versus 73.3 (6.9), p<0.001 for Adelaide students and 72.5 (7.8) versus 70.6 (8.8), p<0.001 for Montreal students. The correlation coefficients were ≥ 0.928 (p<0.001).

Conclusions: Student cohorts performed similarly regardless of which ERP key was used. With appropriate editorial control, SCT's can be effectively shared between French and English-speaking institutions located in different hemispheres. Potential advantages include the creation of an international database of assessment items, benchmarking and cost sharing.

Keywords: script concordance test; international collaboration; bilingual collaboration; medical student assessment;

clinical reasoning

Introduction

Clinical reasoning (CR) is a cornerstone of medical practice. Whilst most methods of teaching and assessing CR have their own benefits and drawbacks a common element is that testing CR is resource-intensive. Thus, the potential for efficiency gains by multi-institutional collaboration in the development of assessment tools is attractive, as is cross-institution benchmarking.

Testing the clinical reasoning of medical students is a core component of assessment in medical programs. Tests of CR are now becoming increasingly important in the postgraduate domain, where it is recognized that errors in CR make the single most significant contribution to successful malpractice claims. (Saber Tehrani 2013) The Script Concordance Test (SCT) is a tool for assessment of CR that is increasingly being used in continuing professional development in medicine (Ahmadi et al 2014) including in large, geographically dispersed medical communities. (Hornos et al 2013)

Script theory explains how physicians progressively acquire knowledge adapted to their clinical tasks. (Charlin et al 2000a, Charlin et al 2000b) The SCT is a written assessment based on clinical scenarios designed to measure clinical data interpretation. 10-20 members of the expert reference panel are recommended for optimal reliability. (Gagnon et al 2005) One significant characteristic of the SCT format is that it allows testing in ill-defined contexts that are often typical of clinical practice. (Lubarsky et al 2013) The SCT has been used in assessment in disciplines including radiology, neurology, radio-oncology, surgery, emergency pediatric medicine, and has been used as an assessment tool for intraoperative decision-making in gynecological surgery. (Brailovsky et al 2001, Brazeau-Lamontagne et al 2004, Lambert et al 2009, Lubarsky et al 2009, Meterissian et al 2006, Park et al 2010) In these reports, tests were statistically reliable and showed construct validity (Lubarsky et al 2011), with statistically linear progression of scores with clinical experience.

These studies in postgraduate medicine have been undertaken with participants of differing levels of clinical expertise. A few studies have assessed reasoning among same level medical students in specific domains. (Collard et al 2009, Duggan 2007, Duggan and Charlin 2012, Monnier et al 2011)

We report our experience in the development and application of a "trans-national, bi-lingual" SCT developed for assessment of senior medical students in Obstetrics and Gynaecology. The research questions were: 1) can SCT questions in Obstetrics and Gynaecology developed in the French-speaking University of Montreal, Canada be readily adapted for use in the English-speaking University of Adelaide, Australia? 2) Could the independent expert reference panels from both institutions be used interchangeably? 3) Would student cohorts in both centres perform to an equivalent level in the same test?

Methods

Background

The University of Montreal has a four-year postgraduate medical program with Obstetrics and Gynaecology taught in a clinical clerkship of 8 week’s duration (4 weeks in Obstetrics and 4 weeks in Gynaecology). Students choose whether to undertake this clerkship in the third or fourth year of their program. In contrast, the University of Adelaide has a six-year undergraduate program with Obstetrics and Gynaecology taught in 9-week clinical clerkships in the fifth year.

Structure, production and scoring of the SCT cases and questions

The SCT format is shown in Figure 1. This provides a clinical scenario (case), a hypothesis or plan of action based on the scenario, and some additional information that may or may not have an effect on the hypothesis or plan. Each scenario is followed by a number of questions. For each question, the participant selects the single best Likert response that describes the effect of the additional information that has been given. In contrast to many conventional forms of written testing (e.g. multiple choice questions), there is no single correct answer; several responses to each question may be considered acceptable. Credit is assigned to each response based on the proportion of experts on the reference panel choosing that response. A maximal score of 1 is given for the response chosen by most of the experts (i.e., the modal response). Other responses are given partial credit in proportion to the number of experts choosing them. (Lubarsky et al 2013)

Figure 1: an example of a SCT case vignette with two questions.

The information in each question stands alone – i.e. when considering the answer to Q1 there is no oxygen saturation result available and for Q2 there is no chest X-ray result available. Typically, between 3-5 questions are provided per clinical case.

You are called to a hospital ward to evaluate a 74-year-old woman three days following vaginal hysterectomy and anterior repair for prolapse. She is complaining of a sore leg and now feels short of breath whilst sitting in a chair.

	If you are considering the following investigation ...	and then you find ...	you would then consider the proposed investigation to be …
Q1	A ventilation-perfusion scan to rule out pulmonary embolism	Her chest X-ray demonstrates areas of collapse	• much less useful • slightly less useful • neither less nor more useful • slightly more useful • much more useful
Q2	An arterial blood gas	Her oxygen saturation whilst breathing room air is 96%	• much less useful • slightly less useful • neither less nor more useful • slightly more useful • much more useful

Development and deployment of questions: Montreal

The Discipline of Obstetrics and Gynaecology in the University of Montreal developed 27 clinical cases in Obstetrics and Gynaecology comprising 82 questions. Faculty members familiar with the Montreal curriculum in

Obstetrics and Gynaecology wrote the questions (in their native French language) taking into account clerkship educational objectives. Questions for the expert reference panel (ERP) were placed on line in a purpose-written restricted access electronic database. The Montreal ERP comprised 15 volunteer experts in Obstetrics and Gynaecology (specialists and subspecialists) who were actively involved in teaching in the university. In Montreal, 154 of 171 (90%) medical students who had completed clinical clerkships in Obstetrics and Gynaecology in four consecutive rotations agreed to sit under normal examination conditions the 82-question paper-based SCT. This part of the study received approval from the University's Institutional Review Board. Neither institution required ethical approval for the remainder of the study. The complete question set of this examination was forwarded electronically to the University of Adelaide.

In brief, the "fate" of the Montreal-derived questions was as follows:

82 questions (27 cases) were received, the first filter removed 5 questions, the second filter removed 8 questions, and the third filter removed 38 questions. The final set comprised 31 questions from 17 cases (range of 1-3 questions per case – see table 1). 31 questions in topics from Obstetrics and Gynecology was the maximal allowed in the assessment blueprint for the clerkship examination in the University of Adelaide.

The first filter (formatting compatibility)

An Adelaide-based, specialist obstetrician and gynecologist (AR) translated the questions in to English. Of the 82 questions 77 were in a suitable format for entry into the Adelaide on-line SCT database. 5 questions related to two cases had Likert anchors of a mixed format (i.e. mixed hypothesis, investigation or management type questions in the one case scenario), which did not fit the configuration of the Adelaide on-line test facility. An Adelaide ERP, made up of 12 Obstetrics and Gynaecology specialists and subspecialists actively involved in teaching, answered these remaining 77 questions on-line.

The second filter (post hoc review of questions for curriculum compatibility, language and transposition errors).

After completion of the Adelaide ERP work a review of the 77 questions was independently undertaken by PD. 8 questions had ambiguous phrases or key errors in translation or transposition of data and were removed. Examples include transposition errors for key laboratory data and omission of a key word such as "only". One question though appropriately translated was considered to be ambiguous - in the case of a woman with severe pre-eclampsia the phrase "managing her conservatively" could have been interpreted by students as meaning observation only or observation plus antihypertensive therapy. This left 69 questions in the question bank and that were suitable for use in the clerkship examinations.

The third filter (selection of 31 questions for use in Adelaide end of year examinations)

The benchmarking exercise to be undertaken was to compare the performance of Montreal and Adelaide student cohorts in identical SCT questions. A representative sample comprising 17 clinical cases and 31 questions was chosen covering diagnosis, management and investigation of common clinical conditions (table 1).

Table 1: Overview of the 18 topics covered in shared questions utilised in assessments in Adelaide and Montreal based medical students (17 clinical cases and 31 questions, 1-3 questions per case).

Question category	Topic

Differential diagnosis	Acute pelvic pain (non-pregnancy related) Pain and bleeding in early pregnancy Postmenopausal bleeding Postpartum haemorrhage Small for dates pregnancy Suspected domestic violence Suspected gestational diabetes mellitus Vulvovaginal irritation
Management	Abnormal menstrual bleeding Analgesia in labour Cervicitis Contraception Intrapartum monitoring of the fetus Pain and bleeding in early pregnancy Urinary incontinence Vaginal prolapse
Investigation	Pre-eclampsia Reduced fetal movements

Question topics were selected with reference to our assessment blueprint for the end of year examination and after applying the Adelaide criteria for selection of SCT questions (Duggan and Charlin 2012) 1) the modal response was consistent with current best evidence 2) alternatives to the modal answer were chosen by more than one member of the ERP 3) the questions reflected the contents of the curriculum 4) the questions were of an appropriate degree of difficulty for the Year 5 cohort 5) the questions complemented our other assessment items.

Statistical analysis

The effect on Adelaide and Montreal student scores and the correlation of scores obtained with the different expert reference panel keys were analysed using the paired samples t-test function in SPSS version 20 for Mac.

Results

There were 123 students (67 female, 56 male) in the fifth year Adelaide cohort and 154 students (98 female, 42 male, 14 no data) in the Montreal cohort. 92 Montreal students reported they were in third year, 37 in fourth year, and 25 did not declare their year.

Mean (SD) percentage scores were 74.2 (6.4) versus 73.3 (6.9), p< 0.001 for the Adelaide students and 72.5 (7.8) versus 70.6 (8.8), p< 0.001 for the Montreal students. Overall, the 95% confidence interval of the difference of the means was 0.4% - 2.4% (table 2).

Table 2. Result of a 31-item SCT paired samples statistical analysis (SPSS 20 for Mac) for Adelaide and Montreal

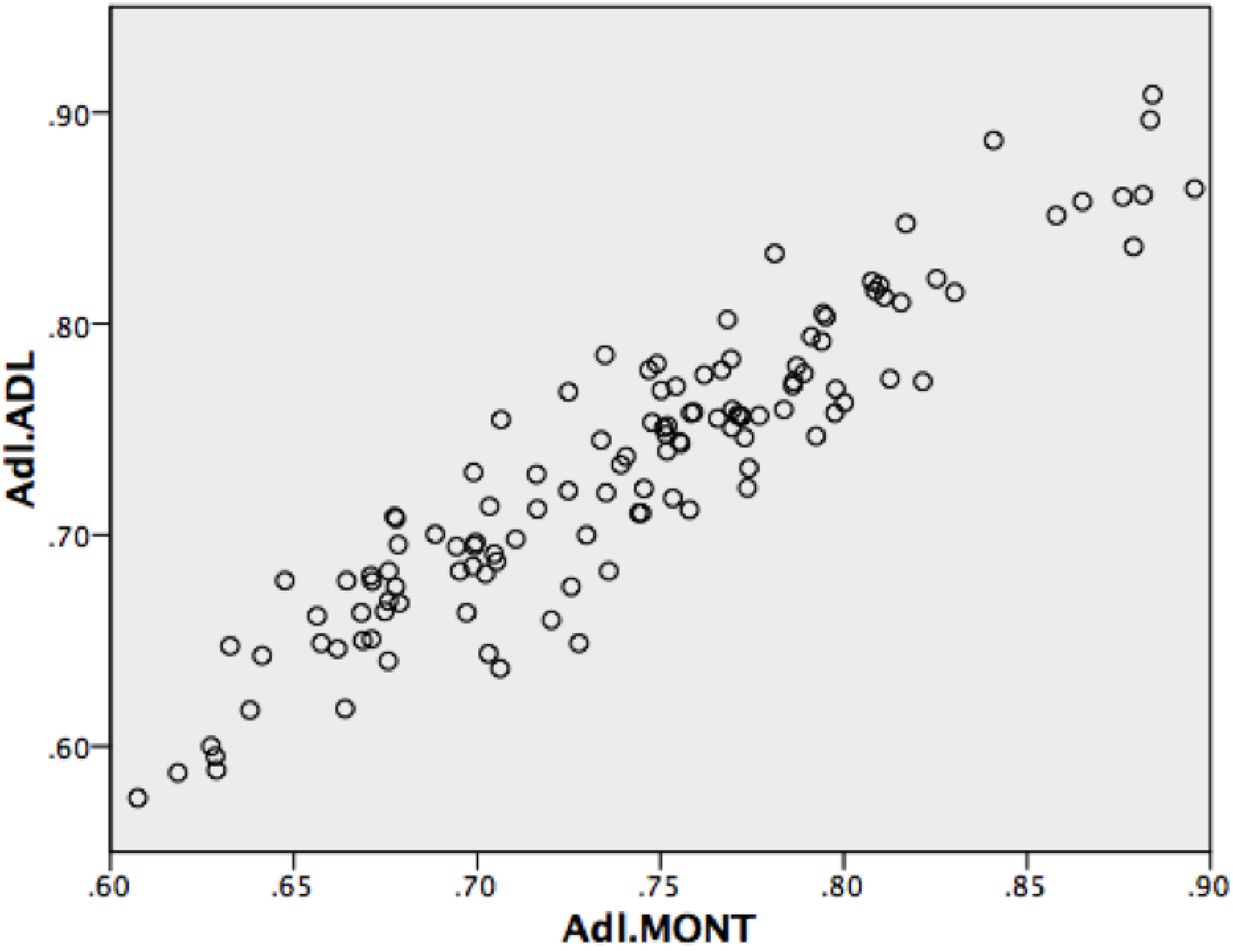
student cohorts as determined by Adelaide-based and Montreal-based expert panels of reference. St. Dev. = standard deviation, ERP = expert reference panel, CI = confidence interval.

Paired Samples Statistics					
Student cohort	Expert reference panel	Mean	n	Std. Dev.	Std. Error Mean
Adelaide	Adelaide ERP	0.733	123	0.070	0.006
	Montreal ERP	0.742	123	0.064	0.006
Montreal	Adelaide ERP	0.706	154	0.088	0.007
	Montreal ERP	0.725	154	0.079	0.006
Paired Samples Correlations					
Student cohort	Expert reference panel	n	r	Sig.	
Adelaide	Adelaide and Montreal ERP's	123	0.929	<0.001	
Montreal	Adelaide and Montreal ERP's	154	0.943	<0.001	

Paired Samples Test		Paired Differences				Sig.
		Mean	Std. Dev.	95% CI of the Difference		(2-tailed)
Student cohort	Expert reference panel			Lower	Upper	
Adelaide	Adelaide and Montreal ERP's	-0.009	0.026	-0.013	-0.004	<0.001
Montreal	Adelaide and Montreal ERP's	-0.019	0.030	-0.024	-0.014	<0.001

The correlation coefficient for the Adelaide student data run with Adelaide and Montreal ERP scoring keys was 0.93 and the scatter plot of those data are shown in figure 2. Very similar results were obtained when the data for the Montreal student cohort were analysed (correlation coefficient 0.94, scatter plot not shown).

Figure 2. Scatter plot diagram of Adelaide student scores for the 31-item SCT run with expert reference panel scoring keys from Adelaide (y axis) and Montreal (x axis), correlation 0.928, p<0.001.



Discussion

To our knowledge this is the first report of an international collaboration in assessment in Obstetrics and Gynaecology, made all the more remarkable for its bilingual nature. We had 3 research questions: 1) can SCT questions in Obstetrics and Gynaecology developed in the French-speaking University of Montreal, Canada be readily adapted for use in the English-speaking University of Adelaide, Australia; 2) could the independent expert reference panels from both institutions be used interchangeably, and 3) would student cohorts in both centers perform to an equivalent level in the same test?

The key finding of our study was that the English-speaking University of Adelaide was able to utilise the majority of questions developed in the French-speaking University of Montreal. Whilst significant editorial input was required, this is also necessary for locally written questions, in any institution. Furthermore, the level of attrition of questions provided by Montreal to Adelaide is similar if not better than the outcome of many local question-writing workshops. The development of examination questions of any sort is a resource intensive process. There is significant potential for cost sharing in addition to enhanced opportunities for professional development of faculties in this arrangement.

In relation to our second research question, although the differences in means for each cohort were statistically significant the absolute difference was small and well in keeping with, if not better than, results when different local expert reference panels are used. The effect of this difference on real student scores is remarkably small, for example, in comparison with differences reported in multi-institutional standard setting studies of other forms of assessments, such as OSCE examinations. (Boursicott et al 2006) The correlation between results achieved by our student cohorts was very high and it is highly unlikely that better correlations would be obtained between two different reference panels from the same institution. These data lend support to the notion that institutions can use in their assessments scoring keys derived from other institutions. The resourcing implications of this finding are potentially significant and especially so in small to medium sized institutions that might have difficulty in local recruitment of an adequate numbers of panelists. Is it appropriate to use a scoring key developed by a different faculty in another country? There are some risks in doing so, particularly if the assessment is high stakes. These risks can be ameliorated but not eliminated by appropriate local editorial control and cross-institutional benchmarking.

Our third research question was answered in the benchmarking exercise. Benchmarking by definition requires sharing of questions. In the exercise we have described, the cohorts were compared to each other rather than an external control group. The questions chosen for this comparison were determined by both faculties to be appropriate to their curricula. To our knowledge such an international benchmarking exercise has not previously been conducted. We believe that the small difference in scores between the two student cohorts, although statistically significant, is consistent with an equivalent and satisfactory level of performance. However, the two cohorts are not directly comparable in part due to differences in structure and length of the two programs. Furthermore, the Montreal students sat their SCT as volunteers (90% participation rate) with no stakes attached, whereas the Adelaide students were sitting a summative test.

Conclusions

Our study clearly demonstrated that SCT questions in Obstetrics and Gynaecology can be effectively shared between French and English speaking institutions located in different hemispheres. ERP data derived from the collaborating institution can be used provided there is appropriate local editorial control. There appeared to be few differences in clinical practice. Potential advantages include the creation of an international database of assessment items, benchmarking and cost sharing. This will save time for teachers and can be a first step for standardisation in assessment, particularly useful in a world where faculty frequently moves from a country to another.

Take Home Messages

- Script Concordance Test questions can be effectively shared between French and English speaking institutions located in different hemispheres.
- International collaboration in question development creates opportunities for benchmarking and cost sharing.
- Scoring keys derived from independent expert reference panels of collaborating institutions can be used interchangeably.

Notes On Contributors

Paul Duggan is Head of the Discipline of Obstetrics and Gynaecology and Chair of the MBBS Assessment Committee of the University of Adelaide, and a Gynaecologist based at the Central Adelaide Local Health Network, Adelaide, Australia.

Patricia Monnier is Associate Professor in the Department of Obstetrics and Gynecology, Division of Reproductive Endocrinology and Infertility, The Research Institute of the McGill University Health Centre, Royal Victoria Hospital, Montreal, Canada.

Alphonse Roex is Senior Lecturer in the Discipline of Obstetrics and Gynaecology, the University of Adelaide and an Obstetrician and Gynaecologist based at the Lyell McEwin Health Service, Elizabeth Vale, Adelaide, Australia.

Maree-Josee Bédard is Head of the Department of Obstetrics and Gynaecology, Hôpital Saint-Luc, Le Centre hospitalier de l'Université de Montréal (CHUM), Montreal, Canada.

Bernard Charlin is Professor and Head of Research and Development in the Centre for Pedagogy Applied to the Health Sciences (CPASS), University of Montreal, Montreal, Canada.

Acknowledgements

Bibliography/References

Ahmadi S-F, Khoshkish S, Soltani-Arabsahi K, Hafezi-Moghadam P, Zahmatkesh G, Heidari P, Baba-Beigloo D, Baradaran HR, Lotfipour S. Challenging script concordance test reference standard by evidence: do judgments by emergency medicine consultants agree with likelihood ratios? International Journal of Emergency Medicine 2014, 7:34
http://dx.doi.org/10.1186/s12245-014-0034-3

Boursicot K, Roberts T and Pell G. Standard Setting for Clinical Competence at Graduation from Medical School: A Comparison of Passing Scores Across Five Medical Schools. Advances in Health Sciences Education 2006 11:173–183.
http://dx.doi.org/10.1007/s10459-005-5291-8

Brailovsky C, Charlin B, CoteÂ S, and 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. Medical Education 2001;35:430±436

Brazeau-Lamontagne L, Charlin B, Gagnon R, Samson L and Van Der Vleuten C. Measurement of perception and interpretation skills during radiology training: utility of the script concordance approach. Medical Teacher, Vol. 26, No. 4, 2004, pp. 326–332. DOI: Charlin B, Tardif J and Boshuizen H. Scripts and Medical Diagnostic Knowledge: Theory and Applications for Clinical Reasoning Instruction and Research. Acad. Med. 2000;75:182–190.

Charlin B, Tardif J and Boshuizen H. Scripts and Medical Diagnostic Knowledge: Theory and Applications for Clinical Reasoning Instruction and Research. Acad. Med. 2000;75:182–190.

http://dx.doi.org/10.1097/00001888-200002000-00020

Charlin B, Roy L, Brailovsky C, Goulet F, van der Vleuten C. The Script Concordance test: a tool to assess the reflectice clinician. Teach Learn Med. 2000 Fall;12(4):189-9
http://dx.doi.org/10.1207/S15328015TLM1204_5

Collard A, Gelaes S, Vanbelle S, Bredart S, Defraigne J, Boniver J and Bourguignon J. Reasoning versus knowledge retention and ascertainment throughout a problem-based learning curriculum. Medical Education 2009: 43: 854–865.
http://dx.doi.org/10.1111/j.1365-2923.2009.03410.x

Duggan P. Development of a Script Concordance Test using an electronic voting system. Ergo 1, 1 December 2007; 35-41

Duggan, Paul and Charlin, Bernard. Summative assessment of 5th year medical students' clinical reasoning by script concordance test: requirements and challenges. BMC Med Ed 2012 12:29.

Gagnon R, Charlin B, Coletti M, Sauve E and 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? Medical Education 2005; 39: 284–291
http://dx.doi.org/10.1111/j.1365-2929.2005.02092.x

Hornos EH, Pleguezuelos EM, Brailovsky CA, Harillo LD, Dory V, Charlin B. The practicum script concordance test: an online continuing professional development format to foster reflection on clinical practice. J Contin Educ Health Prof. 2013 Winter;33(1):59-66.
http://dx.doi.org/10.1002/chp.21166

Lambert C, Gagnon R, Nguyen D, Charlin B. The script concordance test in radiation oncology: validation study of a new tool to assess clinical reasoning. Radiat Oncol. 2009 Feb 9;4:7.
http://dx.doi.org/10.1186/1748-717X-4-7

Lubarsky S, Chalk C, Kazitani D, Gagnon R, Charlin B. The script concordance test: a new tool and assessing clinical judgment in neurology. Can J Neurol Sci. 2009;36:326-31.
http://dx.doi.org/10.1017/S031716710000706X

Lubarsky S, Charlin B, Cook DA, Chalk C, van der Vleuten CPM. Script Concordance Method: A Review of Published Validity Evidence. Medical Education 2011; 45(4):329-38
http://dx.doi.org/10.1111/j.1365-2923.2010.03863.x

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.
http://dx.doi.org/10.3109/0142159X.2013.760036

Meterissian S. A Novel Method of Assessing Clinical Reasoning in Surgical Residents. Surg Innov 2006; 13; 115.
http://dx.doi.org/10.1177/1553350606291042

Monnier P, Bédard M-J, Gagnon R, Charlin B. The relationship between script concordance test scores in an

obstetrics-gynecology rotation and global performance assessments in the Curriculum. International Journal of Medical Education. 2011; 2:3-6
http://dx.doi.org/10.5116/ijme.4d21.bf89

Park AJ, Barber MD, Bent AE et al. Assessment of intraoperative judgment during gynecological surgery using the Script Concordance Test. Am J Obstet Gynecol 2010;203:240.e1-6
http://dx.doi.org/10.1016/j.ajog.2010.04.010

Saber Tehrani AS, Lee H, Mathews SC, Shore A, Makary MA, Pronovost PJ, Newman-Toker DE. 25-Year summary of US malpractice claims for diagnostic errors 1986-2010: an analysis from the National Practitioner Data Bank. BMJ Qual Saf. 2013 Aug;22(8):672-80. doi: 10.1136/bmjqs-2012-001550. Epub 2013 Apr 22.
http://dx.doi.org/10.1136/bmjqs-2012-001550

Appendices

Declarations

The author has declared that there are no conflicts of interest.