Association between candidate total scores and response pattern in script concordance testing of medical students

S. H. Wan[1], P. Duggan[2], E. Tor[1] & J. N. Hudson[2]

Abstract

Introduction: The script concordance test (SCT) aims to test clinical decision making and clinical reasoning. This study is a preliminary attempt to understand an alleged test-taking strategy where students avoid extreme response options, potentially threatening the validity of SCT scores. We investigated whether there is a significant association between the propensity to avoid the extreme response options and candidates' overall SCT scores.

Methods: The SCT scores of 660 clinical-year medical students (six cohorts from 2013–2015) were analysed for a possible association with candidates' response pattern. The proportion of middle range response options was calculated. Propensity to avoid extreme response options is defined as a response pattern with 15% or more of middle-range responses compared to those of the expert reference panel. The distribution for candidates with propensity to avoid the extreme options was further investigated using chi-square statistics for possible association with their overall SCT results.

Results: Fifty-five percent of the students from the lowest quartile, compared to 30% from the top quartile, had shown a propensity to avoid the extreme options. The differences were statistically significant ($p < 0.001$) and were consistent among all six cohorts included in this study.

Conclusions: Students whose SCT scores are in the lowest quartile are more likely to avoid the extreme response options in answering SCT questions. For quality assurance in high stakes summative SCTs, it may be worthwhile to select items with expert reference panel's modal answers covering the full 5-point response options.

Keywords: medical education; script concordance; clinical reasoning; assessment.

1 School of Medicine, University of Notre Dame, Australia
2 School of Medicine, University of Adelaide, Australia

Correspondence
Siu Hong Wan
Head of Basic and Clinical Science Domain
Associate Professor, Assessment
School of Medicine
University of Notre Dame
Australia
Tel: +61 2 8204 4479
Email: michael.wan@nd.edu.au

ISSN 1442-1100

Introduction

The script concordance test (SCT) was introduced in 2000 by Charlin, aiming to assess the higher-order clinical reasoning skills of medical students (Charlin, Roy, Brailovsky, Goulet, & van der Vleuten, 2000). It is a useful assessment tool to test clinical reasoning and data interpretation skills, and has been shown to be valid (Lubarsky, Vleuten, Charlin, Chalk, & Cook, 2011).

The SCT is a written format currently in widespread use internationally to test clinical reasoning in health professional education. In recent years, the SCT has been used in various medical disciplines, such as internal medicine, paediatrics, emergency medicine, neurology, surgery, anaesthesia and radiology (Boulouffe, Doucet, Muschart, Charlin, & Vanpee, 2014; Brazeau-Lamontagne, Charlin, Gagnon, Samson, & van der Vleuten, 2004; Carrière, 2009; Drolet, 2015; Nouh et al., 2012; Tan, Tan, Kandiah, Samarasekera, & Ponnamperuma, 2011). The SCT has also been used to assess other discipline areas where classical written multiple-choice questions (MCQs) or short-answer questions (SAQs) are difficult to develop, for example, in assessing medical ethical principles and professionalism (Foucault, Dubé, Fernandez, Gagnon, & Charlin, 2015; Tsai, Chen, & Lei, 2012). While more traditional assessment formats such as MCQs and SAQs tend to assess students' lower taxonomic orders of thinking, SCT questions can be used to assess a higher order of thinking (Palmer, Duggan, Devitt, & Russell, 2010). Some forms of modified essay questions (MEQs) have been shown to fail to assess higher cognitive skills and have been replaced with a SCT examination (Duggan & Charlin, 2012; Palmer et al., 2010).

The SCT has been shown to be both valid and reliable in several studies, including a country-wide validation study (Dory, Gagnon, Vanpee, & Charlin, 2012; Lubarsky et al., 2011; Nouh et al., 2012; Wan, 2015). The reliability of a 60 to 90-minute examination had a Cronbach alpha of 0.7–0.85 (Nouh et al., 2012; See, Tan, & Lim, 2014). Evidence supporting the construct validity based on the progression of SCT performance related to the clinical experience from undergraduate students to post-graduate fellowship training has also been reported (Ducos et al., 2015; Lambert, Gagnon, Nguyen, & Charlin, 2009; Wan, 2014).

The SCT assessment format has been successfully implemented in undergraduate and graduate-entry medical schools, residency and fellowship training worldwide as well as in nursing schools (Chang et al., 2014; Dawson, Comer, Kossick, & Neubrander, 2014; Duggan & Charlin, 2012; Irfannuddin, 2009; Kow, Walters, Karram, Sarsotti, & Jelovsek, 2014; Nouh et al., 2012; Palmer et al., 2010). In fact, SCT is one of the few currently available assessment tools for clinical reasoning in the written format (Nouh et al., 2012). It can be implemented relatively easily in the paper-based format or online, and the scoring can be done electronically.

In a typical SCT question, candidates are presented with a clinical scenario followed by an additional piece of information. They are then asked for the probability of the suggested diagnosis or the appropriateness of a proposed investigation or management.

ISSN 1442-1100

The descriptors for the response options range from ruling out/contraindicated (-2), less likely/less appropriate (-1), neither less nor more likely/appropriate (0), more likely/appropriate (+1) to definitive diagnosis/absolutely necessary (+2).

This process reflects how practising clinicians retrieve their "illness scripts" or network of previous clinical experience (about similar patient encounters) when faced with uncertainty with diagnosis, investigation or management (Lubarsky et al., 2011; Wan, 2015).

In order to allow the students to choose from the full range of the five response options, "much less likely (-2)" rather than "ruling out the diagnosis" and "much more likely (+2)" rather than "definitive diagnosis" are used in the questions in our school (Wan, 2015). Two sample SCT questions on diagnosis and management are shown in Figure 1.

To score these SCT questions, the student's decision is compared to that of a reference expert clinician panel. Students are able to score marks according to the "concordance" in the decision with the majority of the panel. A partial score is given if the decision concurs with a minority of the panel.

Clinical Scenario A
A 42-year-old women presents to the general practice with a lump in the neck which moves upward on swallowing.

	If you were thinking of …	**and then you find that …**	**this hypothesis becomes …**	
1	Multinodular goitre	The lump is smooth and measures around 3 cm in diameter	**A B C D E** -2 - 1 0 +1 +2	**-2:** much less likely **-1:** less likely **0:** neither more nor less likely **+1:** more likely **+2:** much more likely
2	Follicular carcinoma of the thyroid	A hard lymph node is palpable in the left cervical chain	**A B C D E** -2 - 1 0 +1 +2	
3	Toxic nodular goiter	His pulse rate is 60 bpm and he has no significant weigh loss	**A B C D E** -2 - 1 0 +1 +2	

Clinical Scenario B
A 45-year-old woman with a history of asthma presents with acute shortness of breath. She is afebrile. On examination, there is a diffuse expiratory wheeze.

	If you were thinking of …	**and then you find that …**	**then your plan of action becomes …**	
4	Giving morphine for her distress	Her PO2 is 55 mmHg and her PCO2 is 60 mmHg	**A B C D E** -2 - 1 0 +1 +2	**-2:** much less likely **-1:** less likely **0:** neither more nor less likely **+1:** more likely **+2:** much more likely
5	Giving hydrocortisone intravenously	Her blood glucose is 24.2 mmol/L	**A B C D E** -2 - 1 0 +1 +2	
6	Giving 5 mg salbutamol by nebuliser	Her pulse rate is 130 bpm	**A B C D E** -2 - 1 0 +1 +2	

Figure 1. Sample SCT questions.

ISSN 1442-1100

An example of using a formula to calculate the weighted scores is shown in Table 1.

Table 1
Formula to Calculate the Weighted Scores in the SCT

Response Options	-2	-1	0	+1	+2
Number of clinicians choosing the answer (out of 10)	7	3	0	0	0
Formula	7/7	3/7	0/7	0/7	0/7
Student's score	1	0.43	0	0	0

Recent literature on the SCT highlighted the observation that the SCT format of aggregate partial credit scoring can be subjected to the validity threat of candidates' test-taking strategy of simply avoiding the extreme response options (Lineberry, Kreiter, & Bordage, 2013). This is similar to the response style coaching strategies described in situational judgment tests that could increase the candidates' scores significantly (Cullen, Sackett, & Lievens, 2006; McDaniel, Psotka, Legree, Yost, & Weekley, 2011). Candidates might choose to avoid the extreme response options (-2 or +2) thinking that the probability of these responses being correct would be low, or they might have a lack of confidence in choosing such extreme options.

Aims

In the present study, we investigated whether or not there is a significant association between the propensity to avoid the extreme response options in SCT (-2 or +2) and the overall SCT scores.

Methods

Participants

In 2013–2015, SCT examinations were implemented in our graduate-entry medical school in NSW, Australia. We collected de-identified data from six clinical SCT written examinations undertaken by three successive cohorts of penultimate-year clinical students and three successive cohorts of final-year clinical students (n = 660). A set of 40 SCT items was given in each examination. The reference panels consisted of clinician experts who were actively involved in teaching the students, general practitioners and academics. Scoring of the items was done according to the formula described in Table 1.

Analysis

We have operationalised propensity in "avoiding the extreme response options" as cases where a candidate's proportion of answers in the middle range (-1, 0, +1) for all 40 items in the SCT was 15% higher than that of the reference panel's. For example, if the reference panel's response pattern showed 50% of responses in the middle range (-1, 0, +1) in a SCT, then if a student's response pattern showed 67.5% of the answers chosen were in the middle range (-1, 0, +1), the student would be deemed to be adopting a test-taking strategy in avoiding the extreme options (-2 or +2).

ISSN 1442-1100

De-identified data in the form of candidates' response pattern in individual SCT items, their total SCT scores, as well as the response data from the expert reference panel were collated and analysed. The proportion of responses to SCT items in the middle-range response options (i.e., -1, 0 and +1) for individual candidates were calculated. They were compared with an expert reference panel's responses, to identify cases of avoidance of extreme-response options.

Chi-square test of association between propensity in avoiding extreme options by candidates and their actual performance in SCT, i.e., the quartile where their overall SCT scores were located within the cohort, was analysed using IBM SPSS© package version 23.

Ethics approval was given by the Human Research Ethics Committee of the University of Notre Dame, Australia.

Results

A total of 660 clinical-year students from six cohorts in the school (three from final year and three from the third year in the four-year medical course) sat the SCT examination.

Using a chi-square test of independence to compare the frequency of avoidance of extreme-response options and the quartile of candidates' overall performance in SCT, a significant association was found (χ^2 (3, 660) = 26.29, $p < 0.001$) (Table 2). Candidates whose SCT scores were in the lowest (first) quartile were more likely to avoid the extreme response options (55%) than other students. This was followed by students in the second quartile (45%) and then students in the third quartile (33%). Students whose SCT scores were in the top quartile had the lowest incidence of avoidance of extreme-response options (30%).

Table 2
Chi-square Test of Independence Between Candidates' Avoidance of Extreme Responses and Percentile Rank of Their Overall SCT Performance

Percentile Rank of SCT Scores (pooled data from 2013–2015 cohorts)						**Chi-square test of association**	
Avoidance of Extreme Response Options	25th percentile rank and below *(i.e., lowest 25% of SCT scores in cohort)* *Count (%)*	25th percentile rank to 50th percentile rank *Count (%)*	50th percentile rank to 75th percentile rank *Count (%)*	75th percentile rank and above *(i.e., highest 25% of SCT scores in cohort)* *Count (%)*	Total N	X^2 (df)	*p*
	(n = 165)	(n = 165)	(n = 165)	(n = 165)			
Yes	90 (54.55)	74 (44.85)	55 (33.33)	49 (29.70)	**660**	26.29 (3)	< 0.001
No	75 (45.45)	91 (55.15)	110 (66.67)	116 (70.30)			

ISSN 1442-1100

The aforementioned observation from the chi-square analysis of pooled data from 2013–2015 was also evident in the data within each of the cohorts (2013 to 2015). This is reported in Table 3 and Figure 2.

Table 3
Chi-square Test of Independence Between Candidates' Avoidance of Extreme Responses and Percentile Rank of Their Overall SCT Performance—by Cohort (2013 to 2015)

		Avoidance of extreme Response Options		Chi-square test of association	
		Yes	No	Total *(N)*	χ^2 *(df,N)*
Quartile/ Percentile Rank of SCT Scores— 2013 cohort	*Lowest quartile for SCT scores (i.e., lowest 25% of SCT scores) Count (%)*	14 (26.42)	39 (73.58)	212	χ^2 (3,212) = 8.58 p = 0.035
	2nd quartile for SCT scores (25th percentile rank to 50th percentile rank) Count (%)	6 (11.32)	47 (88.68)		
	3rd quartile for SCT scores (50th percentile rank to 75th percentile rank) Count (%)	4 (7.55)	49 (92.45)		
	Top quartile for SCT scores (i.e., highest 25% of SCT scores) Count (%)	7 (13.21)	46 (86.79)		
Quartile/ Percentile Rank of SCT Scores— 2014 cohort	*Lowest quartile for SCT scores (i.e., lowest 25% of SCT scores) Count (%)*	44 (77.19)	13 (22.81)	228	χ^2 (3,228) = 12.14 p = 0.007
	2nd quartile for SCT scores (25th percentile rank to 50th percentile rank) Count (%)	38 (66.67)	19 (33.33)		
	3rd quartile for SCT scores (50th percentile rank to 75th percentile rank) Count (%)	30 (52.63)	27 (47.37)		
	Top quartile for SCT scores (i.e., highest 25% of SCT scores) Count (%)	28 (49.12)	29 (50.87)		
Quartile/ Percentile Rank of SCT Scores— 2015 cohort	*Lowest quartile for SCT scores (i.e., lowest 25% of SCT scores) Count (%)*	32 (58.18)	23 (41.82)	220	χ^2 (3,220) = 13.45 p = 0.004
	2nd quartile for SCT scores (25th percentile rank to 50th percentile rank) Count (%)	27 (49.09)	28 (50.91)		
	3rd quartile for SCT scores (50th percentile rank to 75th percentile rank) Count (%)	21 (38.18)	34 (61.82)		
	Top quartile for SCT scores (i.e., highest 25% of SCT scores) Count (%)	14 (25.45)	41 (74.55)		

ISSN 1442-1100

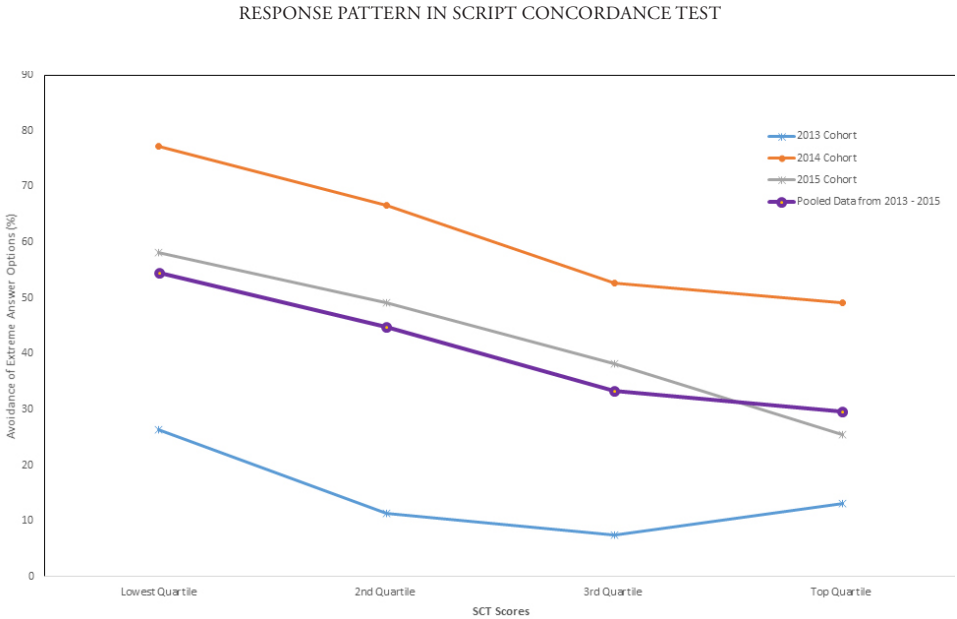

Figure 2. Percentage avoidance of extreme-response options in SCT by candidates' overall performance by quartile in SCT scores.

Discussion

Data from our study shows a significant negative association between overall SCT scores and the propensity to avoid the extreme-response options. This negative association suggests that candidates who tend to avoid extreme-response options do not achieve inflation of their SCT scores, in contrast to the findings from Lineberry, Kreiter and Bordage (2013). A further follow-up study using post-hoc simulation and rescoring of SCT data will provide more evidence on the actual impact of extreme-response options avoidance on candidates' overall SCT scores.

The response pattern, that is, propensity to avoid extreme options, of the students whose SCT scores were in the lowest quartile, could be due to a test-taking strategy or avoidance of the extreme-response options simply because they were not confident about the likelihood of a diagnosis or management plan (due to poor command of basic clinical science knowledge). Such avoidance obviously did not advantage them in terms of getting higher scores.

The aforementioned findings could be a result of some pre-emptive strategies in the study context. Apart from fulfilling the usual blueprinting in ensuring a sufficient spread of clinical scenarios for representativeness of item sampling for each SCT paper, items with roughly equal number of full marks in each option across the five response options are selected from the SCT-item pool. In other words, to mitigate the impact of any test-taking strategies that may have been adopted by students, we select SCT items with modal answers from the expert reference panel that cover the full 5-point

ISSN 1442-1100

Likert scale response options. Students should not be advantaged or disadvantaged by selecting predominantly the "-1", "0" or "+1" response options and avoiding the extreme options of "-2" and "+2". Student performance on SCT tests will then more likely reflect student expertise in clinical reasoning rather than expertise in test-taking behaviour, or confidence in reaching a definitive decision.

While the data for this study only came from one medical school, the study sample was reasonably large (n = 660) and included six cohorts of students. The findings and resulting recommendations related to construction of SCT items should be generalisable to other settings. A limitation of this study is the pure quantitative method used in the analysis. A think-aloud protocol would have been useful to analyse the actual reasons behind the candidates' avoidance of the extreme-response options in SCTs.

Therefore, another study is underway to look at the underlying reasons for candidates avoiding the extreme responses. A focus group discussion and think-aloud analysis will look deeper into what is in the students' mind when they choose to avoid the extreme-response options in SCT, i.e., whether this avoidance behaviour is due to lack of confidence in their command of clinical science knowledge for clinical reasoning and decision making, or it is a conscious test-taking strategy employed by the students.

Before conclusive recommendations can be made, further work to investigate the issue of potential threats to validity of SCT scores are crucial, particularly using empirical data from other medical schools using SCT as an assessment modality. A simulation study through post-hoc rescoring of current SCT data set (as briefly mentioned before) will be conducted in this study context to further investigate the extent of score inflation in SCT as a result of complete avoidance of extreme-response options (by recoding "-2" to "-1"; "+2" to "+1") or as a result of only choosing "0" as the answer to all items which were performed by other colleagues (Lineberry et al., 2013; See et al., 2014).

Conclusions

Students whose SCT scores are in the lowest quartile seem more likely to avoid the extreme-response options in answering SCT questions.

Developing good-quality SCT questions is not easy. As with all other assessment modalities, careful planning and development of SCT items, along with necessary quality assurance and quality monitoring mechanisms, are crucial to mitigate possible threats to the validity of SCT scores. Acknowledging the vulnerability of SCT scores to possible validity threats due to the format of SCT response options and the characteristics of aggregate partial-credit scoring models is crucial. As demonstrated by the study findings, careful construction and selection of items that can be built into the SCT development procedures may be helpful to mitigate some of the plausible threats to validity of SCT scores. Particular care should be taken to develop SCT items that could attract the full range of the 5 response options available for student answer choice. In other words, the additional pieces of new information should result in the consideration of "-2" and "+2" as well as "-1", "0" and "+1" options.

ISSN 1442-1100

References

Boulouffe, C., Doucet, B., Muschart, X., Charlin, B., & Vanpee, D. (2014). Assessing clinical reasoning using a script concordance test with electrocardiogram in an emergency medicine clerkship rotation. *Emergency Medicine Journal*, *31*(4), 313–316. doi:10.1136/emermed-2012-201737

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. *Medical Teacher*, *26*(4), 326–332. doi:10.1080/01421590410001679000

Carrière, B. (2009). Assessing clinical reasoning in pediatric emergency medicine: Validity evidence for a script concordance test. *Annals of Emergency Medicine*, *53*(5), 647–652. doi:10.1016/j.annemergmed.2008.07.024

Chang, T. P., Kessler, D., McAninch, B., Fein, D. M., Scherzer, D. J., Seelbach, E., . . . Education, N. (2014). Script concordance testing: Assessing residents' clinical decision-making skills for infant lumbar punctures. *Academic Medicine*, *89*(1), 128–135. doi:10.1097/ACM.0000000000000059

Charlin, B., Roy, L., Brailovsky, C., Goulet, F., & van der Vleuten, C. (2000). The script concordance test: A tool to assess the reflective clinician. *Teaching and Learning in Medicine*, *12*(4), 189–195. doi:10.1207/S15328015TLM1204_5

Cullen, M. J., Sackett, P. R., & Lievens, F. (2006). Threats to the operational use of situational judgment tests in the college admission process. *International Journal of Selection and Assessment*, *14*(2), 142–155. doi:10.1111/j.1468-2389.2006.00340.x

Dawson, T., Comer, L., Kossick, M. A., & Neubrander, J. (2014). Can script concordance testing be used in nursing education to accurately assess clinical reasoning skills? *Journal of Nursing Education*, *53*(5), 281–286. doi:10.3928/0148434-20140321-03

Dory, V., Gagnon, R., Vanpee, D., & Charlin, B. (2012). How to construct and implement script concordance tests: Insights from a systematic review. *Medical Education*, *46*(6), 552–563. doi:10.1111/j.1365-2923.2011.04211.x

Drolet, P. (2015). Assessing clinical reasoning in anesthesiology: Making the case for the script concordance test. *Anaesthesia Critical Care & Pain Medicine*, *34*(1), 5–7. doi:10.1016/j.accpm.2015.01.003

Ducos, G., Lejus, C., Sztark, F., Nathan, N., Fourcade, O., Tack, I., . . . Minville, V. (2015). The script concordance test in anesthesiology: Validation of a new tool for assessing clinical reasoning. *Anaesthesia Critical Care & Pain Medicine*, *34*(1), 11–15. doi:10.1016/j.accpm.2014.11.001

Duggan, P., & Charlin, B. (2012). Summative assessment of 5th year medical students' clinical reasoning by script concordance test: Requirements and challenges. *BMC Medical Education*, *12*(1), 29–29. doi:10.1186/1472-6920-12-29

Foucault, A., Dubé, S., Fernandez, N., Gagnon, R., & Charlin, B. (2015). Learning medical professionalism with the online concordance-of-judgment learning tool (CJLT): A pilot study. *Medical Teacher*, *37*(10), 955.

ISSN 1442-1100

Irfannuddin, I. (2009). Knowledge and critical thinking skills increase clinical reasoning ability in urogenital disorders: A Universitas Sriwijaya medical faculty experience. *Medical Journal of Indonesia*, *18*(1), 53–59. doi:10.13181/mji.v18i1.341

Kow, N., Walters, M. D., Karram, M. M., Sarsotti, C. J., & Jelovsek, J. E. (2014). Assessing intraoperative judgment using script concordance testing through the gynecology continuum of practice. *Medical Teacher*, *36*(8), 724–729. doi:10.3109/0142159X.2014.910297

Lambert, C., Gagnon, R., Nguyen, D., & Charlin, B. (2009). The script concordance test in radiation oncology: Validation study of a new tool to assess clinical reasoning. *Radiation Oncology*, *4*(1), 7. doi:10.1186/1748-717X-4-7

Lineberry, M., Kreiter, C. D., & Bordage, G. (2013). Threats to validity in the use and interpretation of script concordance test scores. *Medical Education*, *47*(12), 1175–1183. doi:10.1111/medu.12283

Lubarsky, S., van der Vleuten, C. P. M., Charlin, B., Chalk, C., & Cook, D. A. (2011). Script concordance testing: A review of published validity evidence. *Medical Education*, *45*(4), 329–338. doi:10.1111/j.1365-2923.2010.03863.x

McDaniel, M. A., Psotka, J., Legree, P. J., Yost, A. P., & Weekley, J. A. (2011). Toward an understanding of situational judgment item validity and group differences. *The Journal of Applied Psychology*, *96*(2), 327–336. doi:10.1037/a0021983

Nouh, T., Boutros, M., Gagnon, R., Reid, S., Leslie, K., Pace, D., . . . Meterissian, S. H. (2012). The script concordance test as a measure of clinical reasoning: A national validation study. *American Journal of Surgery*, *203*(4), 530–534. doi:10.1016/j.amjsurg.2011.11.006

Palmer, E. J., Duggan, P., Devitt, P. G., & Russell, R. (2010). The modified essay question: Its exit from the exit examination? *Medical Teacher*, *32*(7), e300–e307. doi:10.3109/0142159X.2010.488705

See, K. C., Tan, K. L., & Lim, T. K. (2014). The script concordance test for clinical reasoning: Re-examining its utility and potential weakness. *Medical Education*, *48*(11), 1069–1077. doi:10.1111/medu.12514

Tan, K., Tan, N., Kandiah, N., Samarasekera, D., & Ponnamperuma, G. (2011). A script concordance test for neurological localization and emergencies. *Neurology*, *76*(9), A268–A268.

Tsai, T.-C., Chen, D.-F., & Lei, S.-M. (2012). The ethics script concordance test in assessing ethical reasoning. *Medical Education*, *46*(5), 527–527. doi:10.1111/j.1365-2923.2012.04252.x

Wan, S. H. (2014). Using script concordance testing (SCT) to assess clinical reasoning: The progression from novice to practising general practitioner. *Medical Education*, *48*(2), 6.

Wan, S. H. (2015). Using the script concordance test to assess clinical reasoning skills in undergraduate and postgraduate medicine. *Hong Kong Medical Journal*, *21*(5), 455–461. doi:10.12809/hkmj154572

ISSN 1442-1100