The script concordance test for clinical reasoning: re-examining its utility and potential weakness

Kay C See,[1] Keng L Tan[2] & Tow K Lim[1]

CONTEXT The script concordance test (SCT) assesses clinical reasoning under conditions of uncertainty. Relatively little information exists on *Z*-score (standard deviation [SD]) cut-offs for distinguishing more experienced from less experienced trainees, and whether scores depend on factual knowledge. Additionally, a recent review highlighted the finding that the SCT is potentially weakened by the fact that the mere avoidance of extreme responses may greatly increase test scores.

OBJECTIVES This study was conducted in order to elucidate the best cut-off *Z*-scores, to correlate SCT scores with scores on a separate medical knowledge examination (MKE), and to investigate potential solutions to the weakness of the SCT.

METHODS An analysis of scores on pulmonary and critical care medicine tests undertaken during July and August 2013 was performed. Clinical reasoning was tested using 1-hour SCTs (Question Sets 1 or 2). Medical knowledge was tested using a 3-hour, computer-adapted, multiple-choice question examination.

RESULTS The expert panel was composed of 16 attending physicians. The SCTs were completed by 16 fellows and 10 residents. Fourteen fellows completed the MKE. Test reliability was acceptable for both Question Sets 1 and 2 (Cronbach's alphas of 0.79 and 0.89, respectively). *Z*-scores of − 2.91 and − 1.76 best separated the scores of residents from those of fellows, and the scores of fellows from those of attending physicians, respectively. Scores on the SCT and MKE were poorly correlated. Simply avoiding extreme answers boosted the *Z*-scores of the lowest 10 scorers on both Question Sets 1 and 2 by ≥ 1 SD. Increasing the proportion of questions with extreme modal answers to 50%, and using hypothetical question sets created from Question Set 1 overcame this problem, but consensus scoring did not.

CONCLUSIONS The SCT was able to differentiate between test subjects of varying levels of competence, and results were not associated with medical knowledge. However, the test was vulnerable to responses that intentionally avoided extreme values. Increasing the proportion of questions with extreme modal answers may attenuate the effect of candidates exploiting the test weakness related to extreme responses.

Medical Education 2014: 48: 1069–1077
doi: 10.1111/medu.12514

Discuss ideas arising from the article at www.mededuc.com discuss.

[1]Division of Respiratory and Critical Care Medicine, University Medicine Cluster,National University Hospital, Singapore
[2]Department of Respiratory and Critical Care Medicine, Singapore General Hospital, Singapore

Correspondence: Kay C See, Division of Respiratory and Critical Care Medicine, University Medicine Cluster, National University Hospital, Singapore 119228. Tel: 00 65 9235 9835; E-mail: kay_choong_see@nuhs.edu.sg

INTRODUCTION

Physicians often perform clinical reasoning under conditions of incomplete patient data and uncertainty. As accurate clinical decisions will lead to better quality care and outcomes, the evaluation of physicians' clinical reasoning ability is essential to determine competency for practice.[1] Methods to do this are few, and alternatives such as the long case examination and the objective structured clinical examination are logistically burdensome and difficult to standardise. A less resource-intensive method of assessment is the script concordance test (SCT), developed based on script theory and the hypothetico-deductive clinical reasoning model.[2–4] The SCT is relatively easy to construct, can be machine-scored, and has been psychometrically tested for construct validity and reliability in multiple health science settings, for both clinical and ethical problems, and for individuals across the entire educational continuum.[5–10]

A well-constructed SCT requires careful item development and expert panel selection.[2,11,12] To support the utility of the SCT, we must concurrently justify the use of the test (by determining if the SCT can differentiate among levels of clinical reasoning competence) and demonstrate score validity (by determining if SCT scores reflect variations in participants' clinical reasoning ability rather than in their medical knowledge).[13,14] However, once constructed and administered, there is relatively little information on the *Z*-score (standard deviation [SD]) cut-offs useful for distinguishing more experienced from less experienced trainees,[15–17] and whether scores actually depend on medical (factual) knowledge rather than on clinical reasoning.[2,18,19] Filling these gaps in information would enhance the usefulness of the test for formative or summative assessment, including in assessments utilised for pass/fail decisions. In addition, a recent systematic review highlighted the finding that the SCT is potentially weakened by the fact that the mere avoidance of extreme responses (i.e. responses of + 2 or − 2) may greatly increase test scores.[20] Unless this flaw can be remedied, it may seriously threaten the score validity of the SCT. A possible solution using consensus scoring (i.e. single best-answer scoring) was proposed,[19,21,22] in which only the most popular answer given by a panel of experts would be given a point and any other answers would score nothing.

We thus had a two-fold aim. Firstly, to study the utility of the SCT, we searched for the best cut-off *Z*-scores with which to distinguish junior from senior pulmonary medicine trainees, and correlated test scores for senior trainees with those on a separate medical knowledge examination (MKE). Secondly, to investigate solutions with which to overcome the recently identified weakness of the SCT, we experimented with consensus scoring on a set of real tests. We also hypothesised that a low proportion of questions with extreme modal answers (i.e. questions for which most panel experts selected the + 2 or − 2 responses) would lead to lower overall scores for participants who favoured extreme responses. Hence, we investigated whether an increase in the proportion of questions with extreme modal answers could nullify the test-taking strategy of indiscriminately avoiding extreme test responses.

METHODS

Participants and setting

We performed a prospective analysis of SCTs undertaken by residents and fellows enrolled in training programmes run by two Singaporean hospital medical departments (at the National University Hospital [NUH], Singapore, and at Singapore General Hospital [SGH]) during July and August 2013. Fellows were senior trainees in the first month of a 3-year fellowship programme in pulmonary medicine at NUH and SGH. Residents were recent medicine graduates and junior trainees from NUH in the first month of a 3-year internal medicine residency programme, who were scheduled to be rotated to pulmonary medicine. Both programmes are accredited by the Accreditation Council for Graduate Medical Education International (ACGME-I). Participants also included pulmonary medicine attending physicians at NUH, who served as the panel of experts as required by SCT methodology. All of the attending physicians had practised clinical pulmonology and intensive care medicine for at least 6 years continuously, and had postgraduate specialist qualifications in both fields. Our ethics review board waived the need for informed consent as the tests were undertaken as part of the standard training curriculum (National Healthcare Group Domain-Specific Review Board; F/2013/00779).

Test construction and conduct

The SCT was created during May and June 2013 according to recently released guidelines.[2] A typical SCT item starts with a brief, realistic, but ambiguous clinical scenario (Appendix S1, online). The scenario is followed by one or more questions, each of

which supplies additional information, and asks the participant to rate a suggested response (a diagnostic possibility, investigative option, or therapeutic option) on a 5-point Likert scale ranging from − 2 (strongly against the response) to 0 (neither for nor against the response) to + 2 (strongly for the response). We labelled the Likert scale anchors according to the guideline recommendations and used anchors at the extremes that were neither overly categorical nor unequivocal. This was intended to encourage participants to select options across the range of the Likert scale. A panel of experts taking the SCT will provide a set of answers that reflects the variability of responses in real life. For the answer scheme, full credit is given to modal choices and progressively lower scores are given to less frequently picked choices (i.e. aggregate scoring). No marks are awarded for responses not picked by any of the experts. When non-experts take the SCT, better alignment of participant choices with the more frequent expert panel responses will result in higher overall scores.

The first and senior authors (KCS, TKL), who are both attending physicians in pulmonology and intensive care medicine, generated a bank of scenarios, each with a brief vignette and three questions, based on the American Board of Internal Medicine (ABIM) pulmonary disease blueprint (www.abim.org/pdf/blueprint/pulm_cert.pdf). The scenarios covered clinical reasoning dilemmas in diagnosis, investigation and therapy, in the fields of pulmonary medicine and intensive care medicine. Both authors discussed the questions in detail to ensure clinical relevance and to clarify ambiguous language. The questions spanned a broad range of topics in pulmonary medicine (e.g. pneumonia, tuberculosis, chronic obstructive pulmonary disease) and intensive care medicine (e.g. haemodynamic optimisation, ventilatory failure). We were careful to only include clinically meaningful scenarios, and specifically avoided any areas of excessive controversy or clinical equipoise (e.g. clamping versus no clamping of chest tubes prior to removal in pneumothorax management). As per SCT methodology, we instructed that questions nested within each SCT item should be considered independently of the other questions.

To form the expert panel of attending physicians, we approached the remaining 17 faculty members in the NUH Pulmonary Medicine Programme, all of whom were general respiratory and critical care medicine physicians with a variety of subspecialty interests. Of these, 16 attending physicians agreed to contribute to the test construction, satisfying the optimal number of panel experts.[2,23] These attending physicians separately responded to the questions within a time limit that simulated the test environment, and their answers contributed to the scoring key. Expert panel responses were anonymised and kept confidential. We used the Cronbach's alpha internal consistency coefficient to estimate test reliability and eventually generated a set of 64 scenarios (192 questions). We then split the initial question set into two sets (Question Sets 1 and 2), each containing 32 scenarios (96 questions) to be answered within 1 hour.[12,24] About a third of the scenarios covered each of diagnosis, investigation and therapy, and each question set contained similar proportions of scenarios. We further checked that both Question Sets 1 and 2 covered the key content areas described in the ABIM blueprint.

Fellows and residents were tested separately at the start of their respective training programmes. Instructions on how to answer an SCT using non-pulmonology and non-intensive care-based case examples were e-mailed to all participants several days before the test. Trainees underwent testing in an invigilated, closed-book environment during July and August 2013. Each trainee took either Question Set 1 or Question Set 2. Answers were scored using a freely available Microsoft Excel spreadsheet calculator from the University of Montreal (www.cpass.umontreal.ca/recherche-et-developpement/script-concordance-tests-scts/excel-corrector-program.html). Scores comprised raw marks and *Z*-scores.[15] The latter represented the number of SDs below the mean raw score of the panel of attending physicians (i.e. a *Z*-score of − 2.13 meant that the trainee scored 2.13 SDs below the mean panel score).

In addition, 14 fellows took a 3-hour, computer-adapted, multiple-choice question (MCQ) examination administered by the US Association of Pulmonary and Critical Care Medicine Programme Directors (APCCMPD) (apccmpd.org) for in-coming fellows, which was primarily an evaluation of medical knowledge. Like our SCT, the APCCMPD test content was also developed using the ABIM blueprint. For the APCCMPD examination, we used the scores generated by the online system. These fellows took the SCT first, and then the APCCMPD examination, with a 15-minute break between them. We employed the SCT and the APCCMPD examination to provide individualised trainee feedback (formative assessment).

Statistical analysis

Univariate comparisons of proportions, means and medians were performed using, respectively, Fisher's exact test, Student's *t*-test, and Wilcoxon rank-sum tests. We used the receiver operating characteristic method to determine *Z*-score cut-offs with optimal sensitivity and specificity for the SCT. These score cut-offs were those that best separated the fellows from the attending physicians, and the residents from the fellows. To investigate if clinical reasoning performance was associated with medical knowledge, we plotted the SCT score against the AP-CCMPD examination score and computed the Pearson correlation coefficient.

We conducted a series of experiments to test the effects of intentional avoidance of extreme responses on raw and *Z*-scores. An important effect would be one with a magnitude sufficient to cross the score cut-offs distinguishing more experienced from less experienced participants. To best illustrate this effect, we picked the 10 lowest scores for each experiment and recoded any extreme responses given into less extreme ones, without altering the direction of the response. In other words, we recoded all + 2 responses to + 1, and all − 2 responses to − 1, left neutral (denoted by 0) responses unaltered, and recomputed overall scores. We performed these adjustments firstly for Question Set 1 and Question Set 2. Secondly, we created hypothetical question sets with increasing proportions of extreme modal answers by initially removing some questions with non-extreme modal answers from Question Set 1 and later by replicating the questions with extreme modal answers in Question Set 1. We then performed a final experiment to investigate the effect of consensus scoring on Question Set 1 (i.e. responses scored 1 point only if they were equal to the modal answers and scored no points if they were not). Statistical significance was indicated by a p-value of < 0.05.

RESULTS

The expert panel comprised 16 attending physicians (median age: 39.5 years; interquartile range [IQR]: 34–42 years; six female), who completed questions from both Question Sets 1 and 2. Ten residents (median age: 25.0 years; IQR: 24–25 years; six female) and 16 fellows (median age: 28.5 years; IQR: 27–35 years; 11 female) were randomly tested with either Question Set 1 or Question Set 2. All questions were answered, with no missing data. Test reliability was acceptable for both Question Set 1 and Question Set 2 (Cronbach's alphas of 0.79 and 0.89, respectively). The attending physicians raw scores (and SDs) were similar for both Question Sets 1 and 2, at 75.2 ± 6.2 and 75.7 ± 8.4, respectively, giving a pooled result of 75.5 ± 7.4. A *Z*-score of − 2.91 best separated the scores of residents and fellows (i.e. most residents scored > 2.91 SDs below the mean panel score, and most fellows scored

Table 1 Test results

Results	**Residents** (n = 10)	**Fellows** (n = 16)	**Attending physicians** (n = 32)*
SCT, raw scores, mean ± SD†‡§	54.0 ± 5.9	62.9 ± 7.8	75.5 ± 7.4
SCT *Z*-scores, mean ± SD†‡§	− 3.37 ± 1.01	− 2.11 ± 1.03	NA
MKE, %, mean ± SD	ND	61.4 ± 6.1¶	ND
Best *Z*-score threshold using ROC analysis	− 2.91 (residents versus fellows)	− 1.76 (fellows versus attendings)	
Sensitivity	81.3%	93.8%	
Specificity	70.0%	75.0%	

* n = 32 as each of the 16 attending physicians completed both Question Sets 1 and 2.
† $p < 0.001$ using one-way analysis of variance (ANOVA).
‡ $p < 0.001$ using multiple regression adjusting for Question Set taken.
§ $p < 0.001$ for the comparison between residents and attending physicians, $p < 0.001$ for the comparison between fellows and attending physicians, and $p = 0.005$ for the comparison between residents and fellows.
¶ n = 14 as two fellows did not take the MKE.
MKE = medical knowledge examination; NA = not applicable; ND = not done; ROC = receiver operating characteristic; SCT = script concordance test (maximum raw score: 96); SD = standard deviation.

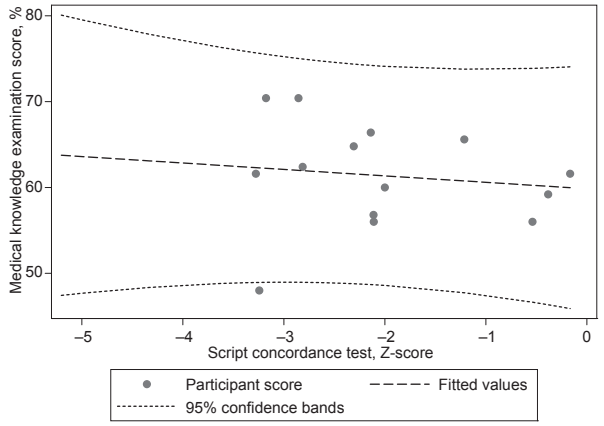

Figure 1 Correlation of scores on the script concordance test with scores on the medical knowledge examination (n = 14 fellows; $r = -0.132$, p = 0.654)

within 2.91 SDs of the mean panel score (Table 1). Similarly, a *Z*-score of − 1.76 best distinguished between the scores of fellows and attending physicians. A scatterplot of SCT scores and the scores on the APCCMPD MKE revealed no obvious trend, which was confirmed by the lack of a significant slope on linear regression analysis (Fig. 1).

For the second part of our study, we computed that only 5% or fewer of our questions in Question Sets 1 and 2 had modal answers that were extreme (i.e. + 2 or − 2). Simply avoiding extreme responses would boost the *Z*-scores of the lowest 10 scorers on either Question Set 1 or Question Set 2 by ≥ 1 SD. This shifted the mean performance of the lowest 10 scorers in Question Set 1 from resident level to fellow level, and improved that of the lowest 10 scorers in Question Set 2 from fellow level to attending physicians level ($p < 0.001$) (Table 2). When we increased the proportion of questions with extreme modal answers, using hypothetical question sets created from Question Set 1, the effect of avoiding extreme answers was attenuated (Question Sets 3–7, Table 2). No significant increase in scores could be seen when the proportion of questions with extreme modal answers reached 50%. Conversely, the use of consensus scoring for Question Set 1 resulted in marked score improvement if participants had avoided extreme responses on purpose (Question Set 8, Table 2)

DISCUSSION

We showed that in test subjects who were naïve to the test strategy, the SCT was able to differentiate among subjects according to level of competence, and results were not associated with medical knowledge. However, the test was vulnerable to a response strategy that intentionally avoided the extreme values. Increasing the proportion of questions with extreme modal answers reduced the effect of subjects exploiting the test's weakness. Rewarding only responses that were exactly equal to the modal answers (consensus scoring) did not appear to solve the problem.

The cut-off *Z*-score that best distinguished among fellows and attending physicians was surprisingly close to the guideline-recommended 2-SD threshold.[2,15] However, no prior studies had evaluated the cut-off *Z*-score that best differentiated residents from fellows. Trainees may theoretically not perform at their best if the examination is framed as a formative rather than summative assessment. However, we were still able to elicit divergent scores for participants with differing levels of competence. We were additionally able to demonstrate reasonably sensitive and specific cut-offs, given that we sampled participant groups who had substantial (≥ 3 years) differences in clinical training and experience. For the 'misclassified' participants, although our modest sample size precluded statistical analysis, we were unable to discern any differences in age, gender and training institution. We believe that other factors may be important determinants of clinical reasoning ability, and these require further investigation.

Although both our SCT and the APCCMPD examination were developed using the same ABIM blueprint, the lack of correlation of clinical reasoning with medical knowledge lends support to the suggestion that these are distinct competencies that probably require to be separately evaluated. The SCT may even be superior to standard MCQs in gauging clinical performance.[19] Furthermore, the SCT applied early in training was able to predict clinical reasoning ability in subjects who were retested 2 years later using two other clinical reasoning tools of known validity (short-answer management problems and simulated office oral tests).[25]

In agreement with a recent critique of the SCT,[20] we found that indiscriminate avoidance of extreme test responses could greatly and significantly increase overall scores, such that clinical reasoning performance improved to the next level of competence. We believe this effect was pronounced as a result of the low proportion of questions with extreme modal answers. The results of our

Table 2 Changes in test results when extreme responses are moderated

Panel answers	QS 1		QS 2		QS 3§		QS 4¶	
Modal answers of + 2 or − 2, n (%)	5/96 (5.2%)		2/96 (2.1%)		5/60 (8.3%)		5/30 (16.7%)	
Mean ± SD raw score of attendings (panel of experts)	75.2 ± 6.2		75.7 ± 9.2		46.1 ± 4.6		23 2 ± 2 2	
Test subject responses	**Raw score**	***Z*-score**	**Raw score**	***Z* score**	**Raw score**	***Z*-score**	**Raw score**	***Z*-score**
Mean ± SD of 10 lowest scores, actual	55.6 ± 5.7	− 3.49 ± 0.88	64.5 ± 3.7	− 2.61 ± 0.53	31.9 ± 4.7	− 3.65 ± 0.78	16.4 ± 2.7	− 3.07 ± 1.24
Mean ± SD of 10 lowest scores, adjusted*	66.4 ± 3.3	− 1.84 ± 0.51	55.8 ± 4.7	− 1.63 ± 0.42	38.7 ± 2.8	− 2.01 ± 0.46	19.4 ± 1.5	− 1.75 ± 0.70
Difference, mean ± SD†	10.7 ± 3.5	1.65 ± 0.54	8.6 ± 3.5	0.98 ± 0.40	6.8 ± 2.5	1.63 ± 0.50	2.9 ± 1.7	1.33 ± 0.79
p-value for difference‡	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001

* Test subjects' responses were recoded thus: +2 as +1; −2 as −1. Original responses between −1 and +1 were not altered.
† Difference between adjusted and actual median scores.
‡ Using the paired t-*test*.
§ Hypothetical QS using QS 1, excluding the first 36 questions with modal answers between −1 and +1.
¶ Hypothetical QS using QS 1, excluding the first 66 questions with modal answers between −1 and +1.
** Hypothetical QS using QS 1, excluding the first 66 questions with modal answers between −1 and +1, and replicating twice more each of the five questions with extreme modal answers.
†† Hypothetical QS using QS 1, excluding the first 66 questions with modal answers between −1 and +1, and replicating four more times each of the five questions with extreme modal answers.
‡‡ Hypothetical QS using QS 1, excluding the first 66 questions with modal answers between −1 and +1, and replicating six more times each of the five questions with extreme modal answers.
§§ Using QS 1, responses scored 1 point only if they were equal to the modal answers and no points if they were not (i.e. consensus scoring).
QS = question set; SD = standard deviation.

hypothetical experiments suggest that increasing the proportion of questions with extreme modal answers could nullify this test-taking strategy. Interestingly, a prior publication by Fournier *et al.*[11] suggested that test answers should be spread over each anchor of the Likert scale, although it gave no advice on the actual distribution of questions with, respectively, extreme and non-extreme answers, and the most recent guidelines do not include this recommendation.[2] Alternatively, the Likert scale descriptors for anchors at the extreme ends of the scale could be moderated. However, doing so may confuse participants by blurring the distinction between intermediate and extreme anchors.

We did not find consensus scoring to be useful. Just as this method of scoring was found to be highly intercorrelated with aggregate scoring,[22] both methods were equally susceptible to manipulation by the avoidance of extreme responses. In addition, others have shown that consensus scoring for the SCT, compared with the usual aggregate scoring method, reduced and hindered expertise detection.[26] Hence, consensus scoring cannot be recommended.

Our study has several strengths. We tested the clinical reasoning test in an SCT-naïve cohort. We were able to show that despite the theoretical vulnerability of the SCT to a test-taking strategy of avoiding

QS 5**		QS 6††		QS 7‡‡		QS 8§§	
15/40 (37.5%)		25/50 (50.0%)		35/60 (58.3%)		5/96 (5.2%)	
31.5 ± 2.7		39.8 ± 3.7		48.1 ± 4.8		52.4 ± 8.7	
Raw score	***Z*-score**	**Raw score**	***Z*-score**	**Raw score**	***Z*-score**	**Raw score**	***Z*-score**
22.4 ± 4.1	− 3.37 ± 1.50	28.3 ± 5.7	− 3.10 ± 1.53	34.3 ± 7.3	− 2.88 ± 1.53	34.6 ± 5.7	− 2.04 ± 0.65
24.5 ± 2.2	− 2.59 ± 0.83	29.7 ± 3.3	− 2.74 ± 0.88	34.8 ± 4.4	− 2.77 ± 0.91	46.3 ± 3.7	− 0.70 ± 0.43
2.1 ± 2.4	0.78 ± 0.91	1.3 ± 3.3	0.36 ± 0.88	0.5 ± 4.1	0.11 ± 0.85	11.7 ± 3.7	1.34 ± 0.43
0.023	0.023	0.230	0.230	0.694	0.694	< 0.001	< 0.001

extreme responses, novices did not exploit this weakness well in real life and scored poorly compared with the expert panel. This allowed subsequent testing of our hypotheses. In addition, we strictly followed existing guidelines in constructing the SCT and incorporated up-to-date recommendations.[2] Furthermore, to show that clinical reasoning is a domain separate from medical knowledge acquisition, we tested candidates using a widely used in-training examination designed by the APCCMPD. The simultaneous conduct of the SCT and the MKE also minimised any potential learning effect of test taking or contamination of results by intervening training.

Our results should be interpreted with the following limitations. Firstly, our sample size was fairly small. Despite this, our study was not underpowered as we managed to obtain moderately sized confidence intervals of the scores, resulting in the clear separation of clinical reasoning performance according to level of competence. Secondly, in the process of testing our hypothesis that a higher proportion of questions with extreme modal answers would bolster the SCT against the test-taking strategy of avoiding extreme responses, we artificially removed 36–66 questions from Question Set 1. However, this did not affect the mean *Z*-score and the confidence intervals greatly. Thirdly, in the testing of the same

hypothesis, we had too few questions with extreme modal answers, and had to resort to replicating the questions and responses in hypothetical question sets. In other words, we had to assume that new questions with extreme modal answers would share the same type of responses. Nonetheless, the eventual result aligns with the statistical expectation that if half of the questions have extreme modal answers, then, by chance alone, the intentional avoidance of extreme responses will result in a 50 : 50 chance of getting the answer wrong: it will be no better than flipping a coin.

Overall, our results lend empirical support to the ability of the SCT to differentiate various levels of clinical reasoning using the aggregate scoring strategy and do not show the consensus scoring method to be superior. The SCT is a valuable tool that tests a specific domain of clinical expertise, distinct from mere factual knowledge recall and interpretation. The value of the SCT is enhanced by its wide applicability, ease of administration and potential for use in test–retest situations.[5] Although we are not aware of any evidence showing that participants intentionally avoid extreme responses, further study is required to formulate an appropriate strategy to strengthen the SCT, such as by trialling real question sets with higher proportions of extreme modal answers. This will be essential before the SCT can be used in high-stakes examinations such as those for graduation or certification purposes.[10] Drawing from this work, we will also construct new question sets with much greater numbers of questions with extreme modal answers.

In conclusion, our findings indicate that the SCT remains a useful method with which to evaluate clinical reasoning. Although we acknowledge that it is vulnerable to the intentional avoidance of extreme responses, increasing the proportion of questions with extreme modal answers seems to offer a solution. If this method of strengthening the SCT can be supported by further research findings, this strategy may become incorporated into guidelines to aid the construction of more robust tests.

Contributors: KCS, KLT and TKL jointly conceived the study and prepared the manuscript. KCS and KLT performed the data extraction. KCS performed the data analysis. TKL supervised the analysis and edited the manuscript. KCS had full access to all data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis. All authors approved the final manuscript for submission.

Acknowledgements: none.

Funding: none.

Conflicts of interest: none.

Ethical approval: requirements for the provision of informed consent were waived by the National Healthcare Group Domain-Specific Review Board as the tests were performed as part of the standard training curriculum (F/2013/00779).

REFERENCES

1 Sniderman AD, LaChapelle KJ, Rachon NA, Furberg CD. The necessity for clinical reasoning in the era of evidence-based medicine. *Mayo Clin Proc* 2013;**88** (10): 1108–14.
2 Lubarsky S, Dory V, Duggan P, Gagnon R, Charlin B. Script concordance testing: from theory to practice: AMEE Guide No. 75. *Med Teach* 2013;**35** (3):184–93.
3 Groves M, Dick ML, McColl G, Bilszta J. Analysing clinical reasoning characteristics using a combined methods approach. *BMC Med Educ* 2013;**13** (1):144.
4 Lubarsky S, Gagnon R, Charlin B. Scoring the script concordance test: not a black and white issue. *Med Educ* 2013;**47**:1159–61.
5 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** (4):189–95.
6 Boulouffe C, Doucet B, Muschart X, Charlin B, Vanpee D. Assessing clinical reasoning using a script concordance test with electrocardiogram in an emergency medicine clerkship rotation. *Emerg Med J* 2014;**31** (4):313–6.
7 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;**33** (1):59–66.
8 Tsai TC, Chen DF, Lei SM. The ethics script concordance test in assessing ethical reasoning. *Med Educ* 2012;**46**:527.
9 Piovezan RD, Custodio O, Cendoroglo MS, Batista NA, Lubarsky S, Charlin B. Assessment of undergraduate clinical reasoning in geriatric medicine: application of a script concordance test. *J Am Geriatr Soc* 2012;**60** (10):1946–50.
10 Nouh T, Boutros M, Gagnon R *et al.* The script concordance test as a measure of clinical reasoning: a national validation study. *Am J Surg* 2012;**203** (4): 530–4.
11 Fournier JP, Demeester A, Charlin B. Script concordance tests: guidelines for construction. *BMC Med Inform Decis Mak* 2008;**8**:18.
12 Dory V, Gagnon R, Vanpee D, Charlin B. How to construct and implement script concordance tests: insights from a systematic review. *Med Educ* 2012;**46**: 552–63.
13 Kane MT. Validating interpretive arguments for licensure and certification examinations. *Eval Health Prof* 1994;**17** (2):133–59; discussion 236–41.

14 Cizek GJ. Defining and distinguishing validity: interpretations of score meaning and justifications of test use. *Psychol Methods* 2012;**17** (1):31–43.
15 Charlin B, Gagnon R, Lubarsky S, Lambert C, Meterissian S, Chalk C, Goudreau J, van der Vleuten C. Assessment in the context of uncertainty using the script concordance test: more meaning for scores. *Teach Learn Med* 2010;**22** (3):180–6.
16 Duggan P, Charlin B. Summative assessment of 5th year medical students' clinical reasoning by script concordance test: requirements and challenges. *BMC Med Educ* 2012;**12**:29.
17 Kania RE, Verillaud B, Tran H, Gagnon R, Kazitani D, Huy PT, Herman P, Charlin B. Online script concordance test for clinical reasoning assessment in otorhinolaryngology: the association between performance and clinical experience. *Arch Otolaryngol Head Neck Surg* 2011;**137** (8):751–5.
18 Lubarsky S, Charlin B, Cook DA, Chalk C, van der Vleuten CPM. Script concordance testing: a review of published validity evidence. *Med Educ* 2011;**45**:329–38.
19 Kelly W, Durning S, Denton G. Comparing a script concordance examination to a multiple-choice examination on a core internal medicine clerkship. *Teach Learn Med* 2012;**24** (3):187–93.
20 Lineberry M, Kreiter CD, Bordage G. Threats to validity in the use and interpretation of script concordance test scores. *Med Educ* 2013;**47**:1175–83.
21 Williams RG, Klamen DL, White CB, Petrusa E, Fincher RM, Whitfield CF, Shatzer JH, McCarty T, Miller BM. Tracking development of clinical reasoning ability across five medical schools using a progress test. *Acad Med* 2011;**86** (9):1148–54.
22 Bland AC, Kreiter CD, Gordon JA. The psychometric properties of five scoring methods applied to the script concordance test. *Acad Med* 2005;**80** (4):395–9.
23 Gagnon R, Charlin B, Coletti M, Sauvé 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.
24 Gagnon R, Charlin B, Lambert C, Carrière B, van der Vleuten C. Script concordance testing: more cases or more questions? *Adv Health Sci Educ Theory Pract* 2009; **14** (3):367–75.
25 Brailovsky C, Charlin B, Beausoleil S, Coté 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.
26 Charlin B, Desaulniers M, Gagnon R, Blouin D, van der Vleuten C. Comparison of an aggregate scoring method with a consensus scoring method in a measure of clinical reasoning capacity. *Teach Learn Med* 2002;**14** (3):150–6.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article:

Appendix S1. Typical script concordance test items with Likert-scale anchors.

Received 15 December 2013; editorial comments to author 26 March 2014, accepted for publication 22 April 2014