How to construct and implement script concordance tests: insights from a systematic review

Valérie Dory,[1,2] Robert Gagnon,[3] Dominique Vanpee[2,4] & Bernard Charlin[3]

CONTEXT Programmes of assessment should measure the various components of clinical competence. Clinical reasoning has been traditionally assessed using written tests and performance-based tests. The script concordance test (SCT) was developed to assess clinical data interpretation skills. A recent review of the literature examined the validity argument concerning the SCT. Our aim was to provide potential users with evidence-based recommendations on how to construct and implement an SCT.

METHODS A systematic review of relevant databases (MEDLINE, ERIC [Education Resources Information Centre], PsycINFO, the Research and Development Resource Base [RDRB, University of Toronto]) and Google Scholar, medical education journals and conference proceedings was conducted for references in English or French. It was supplemented by ancestry searching and by additional references provided by experts.

RESULTS The search yielded 848 references, of which 80 were analysed. Studies suggest that tests with around 100 items (25–30 cases), of which 25% are discarded after item analysis, should provide reliable scores. Panels with 10–20 members are needed to reach adequate precision in terms of estimated reliability. Panellists' responses can be analysed by checking for moderate variability among responses. Studies of alternative scoring methods are inconclusive, but the traditional scoring method is satisfactory. There is little evidence on how best to determine a pass/fail threshold for high-stakes examinations.

CONCLUSIONS Our literature search was broad and included references from medical education journals not indexed in the usual databases, conference abstracts and dissertations. There is good evidence on how to construct and implement an SCT for formative purposes or medium-stakes course evaluations. Further avenues for research include examining the impact of various aspects of SCT construction and implementation on issues such as educational impact, correlations with other assessments, and validity of pass/fail decisions, particularly for high-stakes examinations.

*Medical Education 2012: **46**: 552–563*
doi:10.1111/j.1365-2923.2011.04211.x

Discuss ideas arising from this article at www.mededuc.com 'discuss'

[1]Fonds de la Recherche Scientifique - FNRS
[2]Institute of Health and Society (IRSS), Université catholique de Louvain, Brussels, Belgium
[3]Centre de Pédagogie Appliquée aux Sciences de la Santé, Faculty of Medicine, University of Montreal, Montreal, Quebec, Canada
[4]Emergency Department, Centre Hospitalier Universitaire Mont-Godinne, Université catholique de Louvain, Yvoir, Belgium

Correspondence: Institute of Health and Society, Université catholique de Louvain, Clos Chapelle-aux-champs 30 boîte B1.30.15, 1200 Brussels, Belgium. Tel: 00 32 2 764 3471; Fax: 00 32 2 764 3470; E-mail: valerie.dory@uclouvain.be

INTRODUCTION

Clinical reasoning lies at the heart of medical practice. Research on clinical reasoning has examined the cognitive processes involved and the organised knowledge required.[1] Although many models and theories have been proposed, most current conceptions attempt to bring them together in various forms of dual-process model.[2,3] Dual-process models of reasoning in general propose that reasoning involves both analytic and intuitive (or non-analytic) processes.[2–4] Eva[5] has proposed a relatively simple model in which clinical reasoning is seen as an iterative process involving hypothesis generation and hypothesis testing. Hypothesis generation is most often associated with intuitive reasoning, using pattern recognition, whereas hypothesis testing is generally more analytic, although both types of process may be involved in either stage. Various forms of knowledge organisation have been proposed, including prototypes, instances, causal networks and scripts.[6] Script theory is able to account for both types of process.[7,8] Script activation is usually an intuitive process which leads to hypothesis generation, whereas hypothesis testing is guided by these scripts to focus data collection and allow data interpretation.[7,8]

As clinical reasoning skills represent the cornerstone of medical practice, their development and assessment have, of course, been high on the list of objectives of medical educationalists the world over.[9–11] Instruments designed to measure clinical reasoning skills have followed trends in research on clinical reasoning practice.[9,12] When research focused on uncovering what was then regarded as the supposedly unique process of clinical reasoning, instruments were based on asking examinees to solve a limited number of highly authentic simulations of patient encounters (e.g. patient management problems [PMPs]). Evidence of case specificity and an 'intermediate effect' (in which advanced novices obtain higher scores than experts) led to the abandoning of PMPs. Case specificity led to a redirection of efforts in both research and assessment towards the knowledge possessed by students and practitioners. Examinations using multiple-choice questions (MCQs) provided reliable assessments of relevant knowledge. The reliability and feasibility of MCQs encouraged test developers to extend the range of such tests beyond the testing of factual knowledge to assess the output of clinical reasoning. Multiple-choice questions based on rich descriptions of cases and variations on the format, such as extended-matching items (EMIs),[13] are currently widely used. Critics have pointed out some of their limitations, such as low face validity and concerns regarding the educational consequences of limiting testing to cases that have a straightforward correct answer, fuelling the continued quest for instruments that capture expertise in clinical reasoning.[12] Kreiter and Bergus[10] have suggested that clinical reasoning is something of a 'black box' and that assessing its content requires inference from measuring either its *outputs* (the traditional approach, i.e. examining the results of problem solving) or its *inputs* (a less common approach, i.e. evaluating knowledge organisation and ability to integrate new information). What Kreiter and Bergus[10] describe as a chain leading *forward* from declarative knowledge to knowledge organisation, clinical reasoning and, finally, clinical problem solving, has more often been represented vertically in Miller's pyramid as moving *upwards* from knowledge to application of knowledge, competence and, finally, performance.[14] Some consider clinical reasoning to be an application of knowledge,[15] whereas others view it as a key skill.[1] This ambiguity illustrates the pivotal position of clinical reasoning and the ensuing variety of approaches to its assessment. Although this complexity may lead to some confusion, one current trend in assessment may pave the way to novel solutions to this predicament. Several leading figures have called for integrated programmes of assessment to measure the outcomes of curricula in a multifaceted way.[11,16,17] This conception shifts the focus from somewhat optimistic attempts to develop the single perfect assessment instrument to efforts to design sophisticated programmes in which each instrument complements others and the inevitable weaknesses of each instrument are to some extent compensated for. The quality of instruments should therefore no longer be evaluated in isolation but, rather, in relation to that of others.

Recent developments in the assessment of clinical reasoning include the script concordance test (SCT). Developed at the end of the 1990s by Charlin *et al.*, the SCT is based on script theory and aims to measure clinical data interpretation in ill-defined cases.[18] Both the format and the scoring of SCTs are original. The format presents examinees with an ill-defined case in the form of a brief clinical scenario in which the information provided is insufficient to reach a decision. Each case is then followed by a number of items comprising a lead-in that provides a hypothesis, followed by an additional piece of information. Examinees are asked to evaluate the impact of this new information on the likelihood that the proposed hypothesis is correct (Table 1).

Table 1 Example of a script concordance test case followed by three items

A 25-year-old man presents to your general practice surgery. He has a severe retrosternal chest pain that began the previous night. There is nothing of note in his medical history. He does not smoke. His father, aged 60 years, and his mother, aged 55 years, are both in good health						
If you were thinking of:	And the patient reports or you find upon clinical examination:	This hypothesis becomes:				
Pericarditis	Normal chest auscultation	− 2	− 1	0	+ 1	+ 2
Pneumothorax	Decreased breath sounds in the left chest area with hyper-resonant chest percussion	− 2	− 1	0	+ 1	+ 2
Panic attack	Yellow deposits around the eyelids	− 2	− 1	0	+ 1	+ 2

− 2: ruled out or almost ruled out; − 1: less likely; 0: neither more nor less likely; + 1: more likely; + 2: certain or almost certain

Examinees' responses are compared with the answers of a panel of experts and given credit depending on the number of panel experts who gave the same response. The modal answer is credited with a full point, whereas partial credit is given to other answers provided by panel members (Table 2). Since its original publication nearly 15 years ago, much has been written about the SCT. Indeed, Lubarsky *et al.*[19] recently examined arguments for the validity of the SCT based on 37 published studies. The accumulated evidence on the validity of the SCT in terms of content and internal consistency is strong. Lubarsky *et al.*[19] found some evidence of validity in terms of relationships to other measures (i.e. convergence with other clinical reasoning tests, and divergence from knowledge tests and global competence measures). Furthermore, the SCT is efficient in providing reliable scores in short testing times (around 60–90 minutes). The rationale for using the SCT within programmes of assessment is therefore compelling. However, its sophisticated structure and scoring method have led to many studies on optimal test construction, panel composition and scoring procedures. In view of the wealth and complexity of publications on these practical aspects of the SCT, we propose to synthesise the findings of a systematic review of the literature with the aim of helping users to construct and implement SCTs.

METHODS

We conducted the literature search in February 2011.

Data sources

Databases

We searched MEDLINE, PsycINFO and ERIC (Education Resources Information Centre) using the following search terms: 'script concordance' OR 'script-concordance' OR 'script questionnaire*'. The University of Toronto's Research and Development Resource Base was searched using the queries 'script AND concordance' and 'script questionnaire*'. We also searched Google Scholar using the query 'script concordance'.

Table 2 Example of the habitual computation and answer key to one question based on the responses of a panel of 15 experts. Answers not selected by any of the experts receive no credit. The answer selected by the most experts (modal response) is attributed full credit of 1 point. Other answers are attributed partial credit based on the proportion of experts who selected the answer over the number of experts who selected the modal response (in this case, nine)

Response on a 5-point Likert scale	− 2	− 1	0	+ 1	+ 2
Panel members giving each response, *n*	0	0	2	9	4
Proportional calculation	0/9	0/9	2/9	9/9	4/9
Credit attributed	0	0	0.22	1	0.44

Journals

Academic Medicine, *Advances in Health Sciences Education*, *BMC Medical Education*, *Education for Health*, *Education for Primary Care*, *Medical Education*, *Medical Education Online*, *Medical Teacher* and *Teaching and Learning in Medicine* were searched because of known deficiencies in the indexing of medical education journals.[20]

We also searched the websites of three peer-reviewed journals that are not indexed in MEDLINE, PsycINFO or ERIC; two of these are English-language journals (*International Journal of Medical Education* and *Canadian Medical Education Journal*) and one is a French-language journal (*Pédagogie Médicale*).

Conference proceedings

Proceedings of conferences run by the Association for Medical Education in Europe (AMEE) (2001–2010) and Ottawa conferences (only 2008 available) were searched using the term 'script concordance'. The proceedings of the French-language Société Internationale Francophone d'Éducation Médicale (SIFEM) and Conférence Internationale des Doyens et des Facultés de Médecine d'Expression Française (CIDMEF) (1999–2010) conferences were also searched.

Other sources

Reference lists of retrieved articles were reviewed for further relevant references. ISI Web of Knowledge was used in several ways. We first retrieved publications by the developer of the SCT, Bernard Charlin. Charlin's three most cited articles were identified and citation maps looking for papers citing these three articles were created. The Centre de Pédagogie Appliquée aux Sciences de la Santé (Centre of Pedagogy Applied to the Health Sciences [CPASS], University of Montreal) website hosts a bibliography of articles pertaining to the SCT; this was also reviewed.

Co-authors provided three additional references for two dissertations[21,22] and one article that was then in press.[23]

Inclusion and exclusion criteria

All articles, conference communications and dissertations containing primary data on the SCT were included. Literature reviews, guidelines, editorials, letters and monographs were excluded.

When data from communications or dissertations were published as articles, we ensured that data were consistent and retained only the articles. However, some dissertations provided additional details which were added to the data extracted from the articles.

Data extraction and analysis

We extracted data into an Excel® spreadsheet (Table 3). Data were analysed using Excel® and PASW Version 18 (SPSS, Inc., Chicago, IL, USA).

Table 3 Data extracted from each paper on the script concordance test (SCT) and appraisal criteria

General information	References
	Type of paper: journal article/ conference abstract/dissertation
Purpose of the SCT	Summative assessment
	Formative assessment
	Educational tool
	Outcome measure in an intervention study
Type of research	Case report/justification/ clarification/developmental
Aims of the study	
Description of the SCT	Domain
	Number of cases and items
	Response scale
	Panel composition
	Scoring
Sample	Country in which the study was conducted
	Number of participants
	Stage of learning of participants
	Sampling strategy and response rate
	Representativeness of the sample
Methods	Including psychometric model and statistics used
Results	
Appraisal of methods	Is the design appropriate in view of the study's aims?
	Were sufficient details provided to replicate the study?
Appraisal of results	Credibility
	Originality
	Concordance with other studies

RESULTS

The search yielded 82 relevant references (Fig. S1, online), of which 50 were articles, 27 were conference abstracts, and five were dissertations.[18,21–66] (A full list is available in Appendix S1.)

We were unable to retrieve the full texts of two dissertations despite our efforts to contact the authors. However, data from one of the dissertations have been published at least in part in two papers, which were retrieved. A total of 80 references were analysed.

Script concordance tests have been developed in a variety of health professions and for domains ranging from the very specific (e.g. geriatric incontinence) to the very broad (general medicine) (Table 4). Although the SCT was developed for use at clerkship and residency level, it has also been used for pre-clinical undergraduate students and practitioners (Table 4).

Three-quarters of studies were case reports or pieces of justification research (i.e. 'does it work?') and many of these described test development and provided results on internal consistency and discriminative ability. Only a quarter of studies examined various components of the SCT within a programme of research aimed at exploring how best to construct and implement an SCT (i.e. developmental research),[67] and then only in terms of the impact on internal structure or discriminative ability.

Test construction

We shall focus on the key issues of the numbers of cases and questions to be written. We refer readers to Fournier *et al.*[68] for guidance regarding the content of SCTs.

The evidence regarding the internal consistency of the SCT is strong.[19] However, in most of these studies, it is unclear whether internal consistency was calculated on an item or a case basis. Internal consistency calculations based on items fail to take into account the fact that items nested within the same case cannot be assumed to be completely independent. Such calculations are therefore likely to overestimate internal consistency.

The optimal number of items or cases to reach a 0.8 level of reliability was examined in three studies. One study used the Spearman–Brown formula to determine the number of *cases* required to obtain a reliability of 0.8 (cases had one or two items); the optimum number was 36.[25] Two studies conducted D studies to determine the number of *items* required to obtain a reliability of 0.8, giving results of 130[26] and 48–102.[56] One of these studies specifically set

Table 4 Some characteristics of studies included (80 references were analysed)

Type of publication, *n*		Type of research, *n*		Purpose, *n*		Country, *n*		Language, *n*		Domain, *n*		Stage of learning, *n*	
Journal article	50	Case report	14	Assessment	71	Canada	27	English	55	Internal medicine	18	Pre-clinical undergraduates	19
Conference abstract	27	Justification	46	Outcome in an intervention study	4	USA	18	French	25	General practice/family medicine	10	Clerks	44
Dissertation	3	Developmental	20	Educational tool	5	France	24			General medicine	7	Residents	39
						Other Europe	6			Surgery	9	Practitioners	41
						Other Americas	2			Obstetrics and gynaecology	7		
						Other	7			Paediatrics	5		
										Other medical specialties	13		
										Evidence-based medicine	2		
										Cultural competency	2		
										Other health professions	9		
										Veterinary medicine	1		

out to compare the impact of adding cases versus adding items in cases using generalisability analysis of three SCTs.[56] Adding items rather than cases was more effective in increasing test reliability and would also be preferable in terms of feasibility: more items in fewer cases reduces the workload of test designers and the reading time required of examinees. However, there appeared to be a ceiling effect after three or four items per case. The D studies of the three tests found that 25 cases with three items each would result in tests with G coefficients of 0.75–0.86.[56]

Many studies sought to optimise internal consistency by eliminating items with poor item–total correlations (e.g. item–total correlations of < 0.05). This procedure has led to as many as 70% of items being discarded, although generally the proportion was approximately 25%.[23–34,44,45,64] Test developers should therefore consider submitting examinees to more items (e.g. around 100 items), especially when items have not been used previously and their psychometric properties have not been ascertained in this way. This would represent an estimated testing time of around 90 minutes.

Studies examining the number of cases and items did so through the lens of a test's internal structure. Although internal structure is an important component of validity in its unitary conception,[69] test developers should ascertain that cases and items cover the appropriate content to ensure that the test does indeed provide a valid measure of what it purports to measure.

Panel composition

Using a panel of experts as a benchmark for scoring is an integral part of the SCT. Script concordance tests aim to measure reasoning in ill-defined cases in which experts are likely to differ somewhat in their responses, according to their experiential clinical knowledge. One study examined the ideal number of experts on the panel.[64] The reliability (i.e. internal consistency) of the test in a group of 80 residents scored using a panel of 38 experts was 0.76. Using random samples of panellists to compose smaller panels, the study found that smaller panels yielded less precise estimates of the test's reliability.[64] Samples of at least 10 panellists provided satisfactory estimates of internal consistency and there was little gain when panel sizes exceeded 20. Panel sizes of 10–20 members are therefore required, depending on the stakes involved in the assessment. An unexpected finding was that larger panels led to higher mean scores. This presumably reflects the greater availability of partial credit for uncommon responses. This issue warrants consideration for users who wish to base pass/fail decisions on absolute scores or to compare scores on the same test using panels of different sizes.

Four studies compared panels with different types of expertise.[21,57–59] Two studies looked at differences in terms of level of specialisation. One study in midwifery compared results based on a panel with general expertise in midwifery with results based on a panel with specialised expertise in the various domains tested.[21] Neither examinee scores nor the psychometric qualities of individual items differed significantly. However, the specialised panel yielded less reliable scores ($\alpha = 0.300$ versus $\alpha = 0.395$). A study in dermatology residents found that scores differed mainly when types of practice (hospital versus private practice) rather than specialty (dermatology versus general practice) of panellists were compared.[58]

Two studies looked at differences in terms of the (dis)similarities between the clinical settings of panellists and examinees[59,60]. Although the psychometric properties of scores were comparable, scores were slightly higher when panel members worked in settings similar to those to which examinees had been exposed (i.e. concordance was unsurprisingly higher when panel members represented role models or were similar to role models encountered by examinees). Sibert *et al.*[60] submitted learners from France and Canada to the same SCT and compared their results using two reference panels comprised of experts from France and Canada. Although the discriminative ability of the SCT was similar for scores derived from both panels, mean scores increased when panel members and examinees were from the same country (i.e. there was more concordance). Another study compared scores based on two large panels ($n = 29$) composed of family doctors with and without teaching roles, and found that they were highly correlated (intraclass correlation coefficient: 0.98), although scores based on the panel of family doctors with teaching roles were higher.[59]

Any comparison of outcomes with different panels must consider the fact that each panel introduces a certain amount of error of measurement. Two panels are unlikely to yield exactly the same responses and hence lead to exactly the same examinee scores. Currently, there is no conclusive evidence of systematic differences in terms of internal consistency or discriminative ability depending on the criteria used to select panel members. The higher scores found

when panel experts came from the same 'clinical culture' or were involved in teaching students indicate that care should be taken to select panel experts whose clinical setting is relevant to the test's purpose. For instance, in high-stakes national examinations, panel experts should be selected to adequately cover the spectrum of clinical settings found across the country.

There were no studies examining the effect of training panel experts to answer SCTs. Experts involved in performance evaluations are usually trained to rate examinee performance accurately. By contrast, in SCTs experts are asked to provide their own responses to the test in order to create a benchmark for how examinees *should have* responded.

Examination of the panel's responses

Once panellists' responses have been collected, they can be analysed. Overall disagreement among panel members should be viewed less as measurement error (although this may be partly true) and more as a sign of the ill-defined nature of the cases, which is a key feature of the SCT. Items that elicit a unanimous single answer are likely to be measuring knowledge application in well-defined cases, as traditional case-based MCQs do. However, too much disagreement (such as demonstrated by outlying responses in the opposite direction to those of other panellists) may lead to concerns about the quality of item wording or the expertise of panel members. One study examined the effect of the variability of responses from panel members on the SCT's reliability and discriminative ability.[36] Items were classified using three levels of variability. The sub-test with items of high variability was most discriminating but least reliable. The sub-test with items of low variability was least discriminating and yielded an intermediate effect (i.e. residents outperformed practitioners). The sub-test with items of moderate variability was both discriminating and reliable.

Another study examined the impact of excluding responses considered as 'deviant' either by excluding discordant responses or by altogether discarding responses from discordant panellists.[35] Despite a possible increase in face validity, no psychometric benefit emerged from doing either.

It appears that if an SCT is used to complement traditional MCQ- or EMI-based measures of clinical reasoning, moderate variability in the panel's responses will ensure that the SCT is indeed measuring reasoning in ill-defined cases in a reliable way and in a way that captures expertise (i.e. it discriminates appropriately between levels of training). There have, however, been discussions about using SCTs in place of traditional MCQs and EMIs at earlier stages of training. In this scenario, the SCT might aim to measure clinical reasoning in more clear-cut cases in which less variability would be expected in the panel's responses.

Scoring

The uncommon scoring process of SCTs has led to some controversy. Although initially developed using a 7-point Likert scale, the SCT now uses a 5-point scale. Two studies have compared the use of 5- and 3-point scales for analysis (not for test taking). Results regarding the impact on reliability are contradictory.[26,49] The aggregate scoring method has also led to studies on answer weighting. The traditional scoring method weights alternative responses against the modal answer. Alternatives include weighting against the total number of panel responses or using distance from the modal or mean response. Three studies explored this issue, but used different variants of the original scoring key, which makes it difficult to compare their findings. Results are similar[49,61] or better[26] when using the traditional weighting system in terms of reliability and discriminative ability.

Four studies compared the classic partial credit scoring method with single-best answer scoring, determined by consensus[44] or based on the panel's modal or average response.[26,49,62] The results are inconsistent. One study used item response theory to compare results and found that items displayed similar psychometric properties whichever scoring method was used.[62] Two studies found partial credit scoring to be more reliable[26,49] and one found it to be less reliable.[44] Two studies examined the impact on score range, with contradictory results.[26,44] Two studies examined the effect on discriminative ability, also with contradictory results.[44,49]

One study used a scoring method akin to using kappa coefficients to correct scores for chance agreement. A comparison of this method with the traditional method did not indicate better reliability in either (both were very poor).[63] A similar scoring method was used in another study in which it demonstrated good discriminative ability (but was not compared with the traditional method).[46] Overall, the few studies that did examine the effects of different scoring methods failed to find consistent differ-

ences in terms of internal structure or discriminative ability.

Script concordance test developers suggest using a T-transformation of scores based on a distribution of panellists' scores with a mean of 80 and a standard deviation (SD) of 5 to help examinees and educators interpret test scores.[70] In studies examining the discriminative ability of the SCT, effect sizes between students and experts have been in the range of 0–4.74, typically around 2,[18,22,23,31–55] whereas effect sizes between residents and experts have been in the range of 0–3.2, typically around 1.[18,22,31–43] Effect sizes are in fact SDs (for Cohen's *d*, the pooled SD). Based on the typical values of effect sizes found and putting aside their wide range (and assuming that pooled SDs are similar to SDs of experts' scores), students would be expected to achieve a mean score of around 70 and residents a mean score of around 75. These rough projections should be confirmed in individual situations.

Pass/fail decisions

There were no data regarding actual pass/fail decisions. One study used a receiver operating characteristic (ROC) curve to determine a pass/fail threshold. However, only 71% of experts achieved scores above this threshold.[66] We are aware that some users use the mean of panel members' scores as a benchmark, subtracting 4 SDs to obtain the pass/fail threshold.[71] The composition of the panel appears to have an impact on absolute mean scores and thus may need to be considered.[59,60,64]

DISCUSSION

Our systematic review of the literature on the SCT yielded 82 references, of which 80 were analysed. The SCT was designed to measure skill in clinical data interpretation using experiential knowledge.[18] It has been shown to be a highly feasible instrument with desirable properties in a variety of health professions and diverse clinical specialties. Based on the findings from our extensive review of the literature, the following recommendations can be made regarding test construction, panel composition, analysis of experts' responses and scoring.

Studies indicate that test designers should construct around 25 cases with around 4 items per case in order to allow for the post hoc elimination of poorly performing items (aiming for the inclusion of 75 items). Guidance regarding the content of cases and items is available.[68] Tests containing around 100 items require around 90 minutes of testing time. Testing time could be reduced to around 1 hour once item banks of previously used items with acceptable psychometric properties are constituted, provided this allows sufficient coverage of content to ensure validity.

Once the test has been constructed, experts are recruited and asked to take the test individually. One study has determined that the panel should include 10–20 members, depending on the stakes involved.[64] There are no clear criteria regarding the selection of experts. Script concordance test developers suggest choosing credible practitioners according to the purpose of the assessment.[68]

The answers provided by the panellists should be examined to ascertain the degree to which test cases are ill defined. From this perspective, variability in the panel's responses is a constitutional feature of script concordance testing. Despite the understandable unease generated by evidently deviant responses or answers from panellists who provide consistently deviant responses, test designers should not necessarily seek to eliminate them as there are no psychometric benefits in doing so. Furthermore, variability has been shown to be a key element in the discriminative ability of the SCT and hence its validity. Nonetheless, discussions regarding deviant responses have led some to suggest that consensus procedures might provide 'better' answers (i.e. answers with higher face validity). There is some empirical evidence that a consensus procedure would yield less reliable and less discriminating results. Furthermore, the SCT purports to measure the fit between examinees' scripts and the data provided. Scripts are believed to develop from experience with individual patients, which leads to a significant degree of idiosyncrasy, particularly in ill-defined cases.[7,8] Variability in possible responses is therefore at the core of the argument about the content validity of the SCT in terms of its ability to measure reasoning in ill-defined cases. Idiosyncrasy may nevertheless be limited in well-defined, prototypical cases. Users who wish to test clinical knowledge in less advanced students may therefore want to use the SCT format without using its usual scoring method. However, it is likely (and indeed there is some evidence of this in the form of an intermediate effect[36]) that such tests would be equivalent to traditional vignette-based MCQs or EMIs, which have a long history of successful use in this context. Furthermore, there is some evidence, including from a recent article published after we conducted our literature search, that the habitual

scoring method can be used successfully in less advanced students.[65,72] We certainly recommend using the current aggregate scoring method in programmes of assessment in which the SCT is included as an adjunct to other traditional means of assessment with the aim of measuring clinical reasoning in ill-defined cases.

Panel responses are then used to construct an answer key. The evidence regarding alternative scoring schemes (based on collapsing responses on a 3-point scale or weighting responses against the total number of responses rather than the modal response) is contradictory, but the existing method of scoring has proved satisfactory in all of these studies. There are tools available to compute scores automatically (http://www.cpass.umontreal.ca/sct.html).

Developers of SCTs have suggested standardising scores to facilitate their interpretation by examinees and educators, which is a key element in the educational impact of assessment. They suggest transforming scores based on a distribution of the panellists' scores with a mean of 80 and an SD of 5. Learners would be expected to achieve lower scores than experts.

Norm-referenced standard setting based on the distribution of examinees' scores could be used for assessments that aim to select the best candidates for a limited number of places or for formative purposes or medium-stakes course evaluations. The use of the panel's score as a benchmark for standard setting might also be examined. We concur with Lubarsky *et al.*[19] that much more research is required on this issue in the context of high-stakes summative assessment.

Strengths and limitations

Our literature search included references both in English and in French. This is important in view of the fact that the SCT has been widely used in the French-speaking medical education community, including for actual summative assessment.[30] Indeed, 33% of our references were in French. However, references in other languages were not included. We also sought to obtain data from the grey literature (i.e. according to the Fourth International Conference on Grey Literature: '…that which is produced on all levels of government, academics, business and industry in print and electronic formats, but which is not controlled by commercial publishers' [1997], cited in [20]) by searching non-indexed medical education journals and conference proceedings and requesting experts to provide additional references such as dissertations. However, our search of conference proceedings was limited to those of conferences held by three key organisations in the French- and English-speaking arenas. Lubarsky *et al.*[19] recently published another systematic review focusing on the validity evidence currently available. Although our search yielded more than twice as many relevant references, our findings are largely consistent with those of Lubarsky *et al.*[19]

Although the data selection and extraction processes were conducted in a systematic manner, both were performed by only one researcher. It is usually recommended that these processes are carried out by two people or that agreement on a portion of the data is checked by others.[73] However, given that other members of the present research team are very familiar with the literature on SCTs, significant omissions or misinterpretations of references are unlikely to have occurred in the process of data selection.

Finally, our recommendations are based on the existing literature and can therefore only be as robust as this source. The SCT has not yet been used in large-scale, high-stakes summative assessments. The absence of data should lead potential users in this context to err on the side of caution when implementing these guidelines. Longer tests may, for instance, be required to meet stringent validity and reliability requirements in such circumstances. Furthermore, there has been a relative paucity in systematic programmes of research aimed at clarifying how best to construct and implement an SCT and these studies have limited their evaluation to internal structure or discriminative ability. This reflects the focus of most research on the SCT in general.[19] Although these are major factors in the quality of assessment procedures, further research could be conducted to examine other aspects of SCT validity, such as educational impact, correlations with other assessments and the validity of pass/fail decisions, and the impacts of test construction and implementation on these factors.

CONCLUSIONS

Our systematic review of the literature provides potential users with evidence regarding the optimal implementation of the SCT. In order to obtain reliable scores, test designers should construct around 25 ill-defined cases with around 4 items per case, reflecting clinical practice in the selected domain. They should convene a panel of 10–20

members to sit the test and ensure that most items elicit moderately variable responses from panellists. Although alternative scoring methods have been proposed, the traditional method is both satisfactory in terms of the psychometric properties of scores obtained and congruent with the theoretical basis of the SCT. Suggestions have also been made to facilitate score interpretation for both educators, who must make decisions based on scores, and learners, who require useful feedback from assessment. Further research is needed to examine the impact of test design and implementation on issues other than internal structure and discriminative ability.

Contributors: VD conducted the literature search and extraction of data, participated in the synthesis of findings and was responsible for drafting the paper. RG participated in the conception of the search and extraction strategy and in the synthesis of findings. BC and DV contributed to the synthesis of findings. All authors contributed significantly to the critical revision of the manuscript and approved the final version for publication.
Acknowledgements: none.
Funding: none.
Conflicts of interest: none.
Ethical approval: not applicable.

REFERENCES

1 Norman G. Research in clinical reasoning: past history and current trends. *Med Educ* 2005;**39** (4):418–27.
2 Croskerry P. A universal model of diagnostic reasoning. *Acad Med* 2009;**84** (8):1022–8.
3 Pelaccia T, Tardif J, Triby E, Charlin B. An analysis of clinical reasoning through a recent and comprehensive approach: the dual-process theory. *Med Educ Online* 2011;**16**:5890.
4 De Neys W, Glumicic T. Conflict monitoring in dual process theories of thinking. *Cognition* 2008;**106** (3):1248–99.
5 Eva KW. What every teacher needs to know about clinical reasoning. *Med Educ* 2005;**39** (1):98–106.
6 Custers EJFM, Regehr G, Norman GR. Mental representations of medical diagnostic knowledge: a review. *Acad Med* 1996;**71** (10) (Suppl):55–61.
7 Charlin B, Tardif J, Boshuizen HP. Scripts and medical diagnostic knowledge: theory and applications for clinical reasoning instruction and research. *Acad Med* 2000;**75** (2):182–90.
8 Charlin B, Boshuizen HP, Custers EJ, Feltovich PJ. Scripts and clinical reasoning. *Med Educ* 2007;**41** (12):1178–84.
9 van der Vleuten CP, Newble DI. How can we test clinical reasoning? *Lancet* 1995;**345** (8956):1032–4.
10 Kreiter CD, Bergus G. The validity of performance-based measures of clinical reasoning and alternative approaches. *Med Educ* 2009;**43** (4):320–5.
11 Schuwirth L. Is assessment of clinical reasoning still the Holy Grail? *Med Educ* 2009;**43** (4):298–300.
12 Elstein AS. Beyond multiple-choice questions and essays: the need for a new way to assess clinical competence. *Acad Med* 1993;**68** (4):244–9.
13 Case SM, Swanson DB. Extended-matching items: a practical alternative to free-response questions. *Teach Learn Med* 1993;**5**:107–15.
14 Miller GE. The assessment of clinical skills/competence/performance. *Acad Med* 1990;**9** (Suppl):63–7.
15 Wass V, van der Vleuten C, Shatzer J, Jones R. Assessment of clinical competence. *Lancet* 2001;**357** (9260):945–9.
16 van der Vleuten CP, Schuwirth LW. Assessing professional competence: from methods to programmes. *Med Educ* 2005;**39** (3):309–17.
17 Norcini J, Anderson B, Bollela V *et al.* Criteria for good assessment: consensus statement and recommendations from the Ottawa 2010 Conference. *Med Teach* 2011;**33** (3):206–14.
18 Charlin B, Brailovsky C, Leduc C, Blouin D. The diagnosis script questionnaire: a new tool to assess a specific dimension of clinical competence. *Adv Health Sci Educ Theory Pract* 1998;**3** (1):51–8.
19 Lubarsky S, Charlin B, Cook DA, Chalk C, van der Vleuten CP. Script concordance testing: a review of published validity evidence. *Med Educ* 2011;**45** (4):329–38.
20 Haig A, Dozier M. BEME Guide No. 3: systematic searching for evidence in medical education – Part 1: Sources of information. *Med Teach* 2003;**25** (4):352–63.
21 Gantelet M. *Impact du panel de référence sur les résultats d'un test de concordance de script (TCS) développé en formation initiale des sages-femmes.* Dissertation for the completion of a diploma of the Ecole des cadres sages-femmes. Dijon: Ecole de Cadres Sages-Femmes 2008.
22 Lamia B, Sitbon O. *Evaluation du raisonnement et de la compétence clinique par le test de concordance de script Universités Paris V, VI, XI, XII.* Inter-university diploma dissertation. Paris: Universités Paris V, VI, XI, XII 2006.
23 Deschênes MF, Charlin B, Gagnon R, Goudreau J. Use of a script concordance test to assess development of clinical reasoning in nursing students. *J Nurs Educ* 2011;**50** (7):381–7.
24 Cohen LJ, Fitzgerald SG, Lane S, Boninger ML, Minkel J, McCue M. Validation of the seating and mobility script concordance test. *Assist Technol* 2009;**21** (1):47–56.
25 Carriere B, Gagnon R, Charlin B, Downing S, Bordage G. Assessing clinical reasoning in pediatric emergency medicine: validity evidence for a script concordance test. *Ann Emerg Med* 2009;**53** (5):647–52.
26 Ramaekers S, Kremer W, Pilot A, van Beukelen P, van Keulen H. Assessment of competence in clinical reasoning and decision making under uncertainty: the script concordance test method. *Assess Eval High Educ* 2010;**35** (6):661–73.

27 Goulet F, Jacques A, Gagnon R, Charlin B, Shabah A. Poorly performing physicians: does the script concordance test detect bad clinical reasoning? *J Contin Educ Health Prof* 2010;**30** (3):161–6.

28 Meterissian S, Zabolotny B, Gagnon R, Charlin B. Is the script concordance test a valid instrument for assessment of intraoperative decision-making skills? *Am J Surg* 2007;**193** (2):248–51.

29 Monnier P, Bédard M, Gagnon R, Charlin B. The relationship between script concordance test scores in an obstetrics-gynecology rotation and global performance assessments in the curriculum. *Int J Med Educ* 2011;**2**:3–6.

30 Boulouffe C, Charlin B, Vanpee D. Evaluation of clinical reasoning in basic emergencies using a script concordance test. *Am J Pharm Educ* 2010;**74** (10): 194.

31 Ruiz JG, Tunuguntla R, Charlin B, Ouslander JG, Symes SN, Gagnon R, Phancao F, Roos BA. The script concordance test as a measure of clinical reasoning skills in geriatric urinary incontinence. *J Am Geriatr Soc* 2010;**58** (11):2178–84.

32 Lubarsky S, Chalk C, Kazitani D, Gagnon R, Charlin B. The script concordance test: a new tool assessing clinical judgement in neurology. *Can J Neurol Sci* 2009;**36** (3):326–31.

33 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;**4**:7.

34 Tunuguntla R, Ouslander JG, Symes S, Phancao F, Charlin B, Gagnon R, Roos BA, Ruiz JG. A script concordance test (SCT) to measure clinical reasoning for managing geriatric urinary incontinence (UI). *J Gen Intern Med* 2009;**24**(Suppl 1):10.

35 Gagnon R, Lubarsky S, Lambert C, Charlin B. Optimization of answer keys for script concordance testing: should we exclude deviant panellists, deviant responses, or neither? *Adv Health Sci Educ Theory Pract* 2011;**16** (5):601–8.

36 Charlin B, Gagnon R, Pelletier J, Coletti M, Abi-Rizk G, Nasr C, Sauve E, van der Vleuten C. Assessment of clinical reasoning in the context of uncertainty: the effect of variability within the reference panel. *Med Educ* 2006;**40** (9):848–54.

37 Marie I, Sibert L, Roussel F, Hellot MF, Lechevallier J, Weber J. The script concordance test: a new evaluation method of both clinical reasoning and skills in internal medicine. *Rev Med Interne* 2005;**26** (6):501–7.

38 Sibert L, Charlin B, Corcos J, Gagnon R, Lechevallier J, Grise P. Assessment of clinical reasoning competence in urology with the script concordance test: an exploratory study across two sites from different countries. *Eur Urol* 2002;**41** (3):227–33.

39 Gibot S, Bollaert PE. Le test de concordance de script comme outil d'évaluation formative en réanimation médicale. *Pédagogie Médicale* 2008;**9** (1):7–18.

40 Fournier JP, Thiercelin D, Pulcini C, Alunni-Perret V, Gilbert E, Minguet JM, Bertrand F. Évaluation du raisonnement clinique en médecine d'urgence: les tests de concordance des scripts décèlent mieux l'expérience clinique que les questions à choix multiples à contexte riche. *Pédagogie Médicale* 2006;**7** (1):20–30.

41 Charlin B, Brailovsky C, Brazeau-Lamontagne L, Samson L, Leduc C, van der Vleuten C. Script questionnaires: their use for assessment of diagnostic knowledge in radiology. *Med Teach* 1998;**20** (6):567–71.

42 Sibert L, Giorgi R, Dahamna B, Doucet J, Charlin B, Darmoni SJ. Is a web-based concordance test feasible to assess therapeutic decision-making skills in a French context? *Med Teach* 2009;**31** (4):162–8.

43 Lemay JF, Donnon T, Charlin B. The reliability and validity of a paediatric script concordance test with medical students, paediatric residents and experienced paediatricians. *Can Med Educ J* 2010;**1** (2):89–95.

44 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.

45 Fournier JP, Staccini P, Ladner J, Roussel F, Gagnon R, Charlin B. Utilisation du test de concordance des scripts (TCS) pour la correction de l'épreuve de lecture critique d'article (LCA). *Pédagogie Médicale* 2009;**10** (Suppl 1):40.

46 Collard A, Gelaes S, Vanbelle S, Bredart S, Defraigne JO, Boniver J, Bourguignon JP. Reasoning versus knowledge retention and ascertainment throughout a problem-based learning curriculum. *Med Educ* 2009;**43** (9):854–65.

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

48 Gagnon R, Charlin B, Roy L, St-Martin M, Sauve E, Boshuizen HP, van der Vleuten C. The cognitive validity of the script concordance test: a processing time study. *Teach Learn Med* 2006;**18** (1):22–7.

49 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.

50 Sibert L, Charlin B, Gagnon R, Corcos J, Khalaf A, Grise P. Evaluation of clinical reasoning in urology: contribution of the script concordance test. *Prog Urol* 2001;**11** (6):1213–9.

51 Williams RG, Klamen DL, Hoffman RM. Medical student acquisition of clinical working knowledge. *Teach Learn Med* 2008;**20**(1):5–10.

52 Joly L, Braun M, Fournier JP, Benetos A. Test de concordance de script et apprentissage du raisonnement clinique en gériatrie: intérêt dans la formation en stage des étudiants hospitaliers. *Pédagogie Médicale* 2009;**10** (Suppl 1):S39.

53 Joly L, Braun M, Benetos A. Apprentissage du raisonnement clinique et test de concordance de script en gériatrie: intérêt dans la formation en stage des étudiants hospitaliers. *Rev Med Interne* 2009;**30** (Suppl 2):135–6.

54 Khonputsa P, Besinque K, Fisher D, Gong WC. Use of script concordance test to assess pharmaceutical diabetic care: a pilot study in Thailand. *Med Teach* 2006;**28** (6):570–3.
55 Demeester A. *Evaluation du raisonnement clinique des étudiants sages-femmes par le test de concordance de script*. Master's dissertation. Paris: Université Paris 13 2004.
56 Gagnon R, Charlin B, Lambert C, Carriere B, van der Vleuten C. Script concordance testing: more cases or more questions? *Adv Health Sci Educ Theory Pract* 2009;**14** (3):367–75.
57 Bertrand C, Jabre P, Lecarpentier E, Margenet A, Combes X, Carriere B, Gagnon R, Charlin B, Claudepierre P, Farcet JP. Étude exploratoire du test de concordance de script pour l'évaluation certificative des médecins en télémédecine. *Pédagogie Médicale* 2009;**10** (Suppl 1):42.
58 Bursztejn AC, Cuny JF, Adam JL, Sido L, Schmutz JL, de Korwin JD, Latarche C, Braun M, Barbaud A. Test de concordance de script en dermatologie: évaluation du choix du panel d'experts. *Pédagogie Médicale* 2010;**11** (Suppl 1):69.
59 Charlin B, Gagnon R, Sauve E, Coletti M. Composition of the panel of reference for concordance tests: do teaching functions have an impact on examinees' ranks and absolute scores? *Med Teach* 2007;**29** (1):49–53.
60 Sibert L, Charlin B, Corcos J, Gagnon R, Grise P, van der Vleuten C. Stability of clinical reasoning assessment results with the script concordance test across two different linguistic, cultural and learning environments. *Med Teach* 2002;**24** (5):522–7.
61 Charlin B, Gagnon R, Carriere B, Lambert C. *The Script Concordance Test as a Tool for Assessment in Context of Uncertainty: a Scoring Process Study*. Dundee: Association for Medical Education in Europe 2006.
62 Kreiter CD, Bland AC, Gordon JA. *Comparing Two Methods of Scoring a Script Concordance Test*. Central Group on Educational Affairs, Association of American Medical Colleges, Spring Conference, 7-10th April 2005, Madison, WI.
63 Vanbelle S, Massart V, Giet D, Albert A. Test de concordance de script: un nouveau mode d'établissement des scores limitant l'effet du hasard. *Pédagogie Médicale* 2007;**8** (2):71–81.
64 Gagnon R, Charlin B, Coletti M, Sauve E, van der Vleuten C. Assessment in the context of uncertainty: how many members are needed on the panel of reference of a script concordance test? *Med Educ* 2005;**39** (3): 284–91.
65 Holloway R, Nesbit K, Bordley D, Noyes K. Teaching and evaluating first and second year medical students' practice of evidence-based medicine. *Med Educ* 2004;**38** (8):868–78.
66 Park AJ, Barber MD, Bent AE, Dooley YT, Dancz C, Sutkin G, Jelovsek JE. Assessment of intraoperative judgement during gynecologic surgery using the script concordance test. *Am J Obstet Gynecol* 2010;**203** (3):240–6.
67 Schuwirth L, Colliver J, Gruppen L *et al*. Research in assessment: consensus statement and recommendations from the Ottawa 2010 Conference. *Med Teach* 2011;**33** (3):224–33.
68 Fournier JP, Demeester A, Charlin B. Script concordance tests: guidelines for construction. *BMC Med Inform Decis Mak* 2008;**8**:18.
69 Downing SM. Validity: on meaningful interpretation of assessment data. *Med Educ* 2003;**37** (9):830–7.
70 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.
71 Duggan P, Charlin B. Summative assessment of 5th year medical students' clinical reasoning by script concordance test: requirements and challenges. *BMC Med Educ* (accepted).
72 Humbert AJ, Johnson MT, Miech E, Friedberg F, Grackin JA, Seidman PA. Assessment of clinical reasoning: a script concordance test designed for preclinical medical students. *Med Teach* 2011;**6**:472–7.
73 Hammick M, Dornan T, Steinert Y. Conducting a best evidence systematic review. Part 1: From idea to data coding. BEME Guide No. 13. *Med Teach* 2010;**32** (1): 3–15.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article.

Figure S1. Flowchart of the literature search and selection of references.

Appendix S1. Reference list.

Please note: Wiley-Blackwell is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than for missing material) should be directed to the corresponding author for the article.

Received 1 July 2011; editorial comments to authors 7 September 2011, 8 November 2011; accepted for publication 23 November 2011