Script concordance testing: a review of published validity evidence

Stuart Lubarsky,[1,2] Bernard Charlin,[3] David A Cook,[4] Colin Chalk[1,2] & Cees P M van der Vleuten[5]

CONTEXT Script concordance test (SCT) scores are intended to reflect respondents' competence in interpreting clinical data under conditions of uncertainty. The validity of inferences based on SCT scores has not been rigorously established.

OBJECTIVES This study was conducted in order to develop a structured validity argument for the interpretation of test scores derived through use of the script concordance method.

METHODS We searched the PubMed, EMBASE and PsycINFO databases for articles pertaining to script concordance testing. We then reviewed these articles to evaluate the construct validity of the script concordance method, following an established approach for analysing validity data from five categories: content; response process; internal structure; relations to other variables, and consequences.

RESULTS Content evidence derives from clear guidelines for the creation of authentic, ill-defined scenarios. High internal consistency reliability supports the internal structure of SCT scores. As might be expected, SCT scores correlate poorly with assessments of pure factual knowledge, in which correlations for more advanced learners are lower. The validity of SCT scores is weakly supported by evidence pertaining to examinee response processes and educational consequences.

CONCLUSIONS Published research generally supports the use of SCT to assess the interpretation of clinical data under conditions of uncertainty, although specifics of the validity argument vary and require verification in different contexts and for particular SCTs. Our review identifies potential areas of further validity inquiry in all five categories of evidence. In particular, future SCT research might explore the impact of the script concordance method on teaching and learning, and examine how SCTs integrate with other assessment methods within comprehensive assessment programmes.

*Medical Education 2011: **45**: 329–338*
doi:10.1111/j.1365-2923.2010.03863.x

[1]Department of Neurology and Neurosurgery, McGill University, Montreal, Quebec, Canada
[2]Centre for Medical Education, McGill University, Montreal, Quebec, Canada
[3]Centre for Applied Teaching in Health Sciences (Centre de Pédagogie Appliquée aux Sciences de la Santé [CPASS]), University of Montreal, Montreal, Quebec, Canada
[4]Department of Internal Medicine, Mayo Clinic College of Medicine, Rochester, Minnesota, USA
[5]Department of Educational Research and Development, Maastricht University, Maastricht, the Netherlands

Correspondence: Stuart Lubarsky, Department of Neurology and Neurosurgery and Centre for Medical Education, McGill University, Montreal, Quebec H3G 1A4, Canada. Tel: 00 1 514 934 8060; Fax: 00 1 514 934 8265; E-mail: stuart.lubarsky@mcgill.ca

INTRODUCTION

The script concordance test (SCT) is an assessment instrument originally developed for use in medical education.[1] Over the last 10 years, research into the theoretical underpinnings and psychometric properties of script concordance has accumulated. The SCT has garnered interest for use in a wide and disparate array of health-related fields.[2–6] In neurology, a call for a national initiative to promote script concordance assessment in undergraduate and postgraduate training programmes in Canada was recently sounded.[7]

'Validity' refers to the extent to which the results of an assessment, such as an SCT, accurately reflect desired conclusions (inferences or interpretations).[8] Validity evidence for a given instrument's scores derives from data systematically collected and analysed to support or challenge intended interpretations.[9,10] Validity can never be 'proven'. Rather, sufficient evidence is gathered for a specific context and purpose (e.g. high-stakes examinations, formative assessments, maintenance of certification) until conclusions seem appropriately justified (or not).

Although validity is clearly construct- and context-dependent, some inferences about an assessment's scores may transcend specific educational settings. It may therefore be useful to identify evidence related to a general method of assessment that can be translated with confidence to specific instantiations of the method. For example, multiple-choice question (MCQ) assessment scores are generally accepted as valid for assessing knowledge, provided the assessment is sufficiently long and appropriate development standards have been followed.[11] In low-stakes assessment this argument alone may be sufficient, although for moderate and high-stakes settings the validity of scores would have to be confirmed for the specific assessment and context.

The validity of interpretations of SCT scores has not been established in a systematic way. Our purpose, in the wake of the increasing popularity of SCTs, is to develop a structured validity argument for the interpretation of test scores derived through use of the script concordance method.

THE SCRIPT CONCORDANCE METHOD

Stimulus and response format

Script concordance tests are comprised of a series of short clinical scenarios (cases), each followed by a set of test questions consisting of three parts.[12] For each question, the first part ('If you were thinking of…') provides a hypothesis in the form of a diagnostic possibility, an investigative option, a therapeutic alternative, or a prognostic or bioethical consideration. The second part ('…and then you find…') presents new information, such as a physical examination sign, a pre-existing condition, an imaging study or a laboratory test result, that may (or may not) have an effect on the given option. The question is answered in the third part ('…this hypothesis becomes:'), which contains a 5-point Likert-type response scale (ranging from − 2 to + 2). Examinees indicate on this scale the effect they think the new information (part 2) is likely to have on the proposed hypothesis (part 1). Examples of SCT questions are provided in Table 1.

Scoring system

By contrast with many conventional forms of testing, there are no single best answers to SCT questions; several responses to each question may be considered acceptable. The examinee's response to each question is compared with those of an expert panel. Credit is assigned to each response based on how many of the experts on the panel choose that response. A maximum score of 1 is given for the response chosen by most of the experts (i.e. the modal response). Other responses are given partial credit, depending on the fraction of experts choosing them. Responses not selected by experts receive a score of 0. An example of the SCT scoring system is shown in Table 2.

BUILDING A VALIDITY ARGUMENT

Construct identification

The first step in any validity evaluation entails an identification of the intended construct.[13] The essential purpose of construct identification is to justify a particular interpretation of a test score by explaining the behaviour that the test score summarises.[14] According to its originators, SCT scores are meant to reflect 'a specific skill of clinical competence: the ability to weigh clinical information in light of entertained hypotheses'.[1] The ability to appropriately interpret clinical data, particularly under conditions of ambiguity or uncertainty, is an integral part of the clinical reasoning process[15] and lies at the heart of what some refer to as 'clinical judgement'.[16,17]

Table 1 Two examples of script concordance cases with three questions each

Case 1. You are evaluating a 75-year-old man with right hemiparesis in the emergency room						
If you are thinking of:	And then you find:	Your hypothesis becomes:*				
1 Cerebral abscess	The patient had an ear infection 10 days ago	– 2	– 1	0	+ 1	+ 2
2 Ischaemic stroke	Sudden onset 2 hours ago	– 2	– 1	0	+ 1	+ 2
3 Cerebral metastasis	Normal contrast-enhanced CT head scan	– 2	– 1	0	+ 1	+ 2
Case 2. A patient with diplopia and unilateral ptosis is referred for neurological evaluation						
If you are considering ordering:	And then you find:	Your planned management becomes:†				
1 CT chest scan	The patient is 60 years old	– 2	– 1	0	+ 1	+ 2
2 Anti-acetylcholine receptor antibodies	The patient is in her first trimester of pregnancy	– 2	– 1	0	+ 1	+ 2
3 MRI brain scan	Normal pupils	– 2	– 1	0	+ 1	+ 2

* Case 1: – 2 = ruled out or almost ruled out; – 1 = less probable; 0 = neither more nor less probable; + 1 = more probable; + 2 = certain or almost certain
† Case 2: – 2 = contra-indicated or almost contra-indicated; – 1 = less indicated; 0 = neither more nor less indicated; + 1 = more indicated; + 2 = absolutely or almost absolutely indicated
CT = computed tomography; MRI = magnetic resonance imaging

Categories of evidence

The next step in a structured validity inquiry is to investigate the extent to which assessment scores can be presumed to reflect the intended construct. To conduct this step, we searched the PubMed, EMBASE and PsycINFO databases for peer-reviewed, English- and French-language articles relating to the theoretical underpinnings, construction procedures and psychometric properties of SCTs. Using the combined search terms 'script' and 'concordance', we identified 37 relevant articles. We then reviewed these articles to evaluate the construct validity of the script concordance method, following an established approach for analysing validity data from five categories: content; response process; internal structure; relations to other variables, and consequences.[18]

Content

This first category of validity evidence evaluates 'the relationship between a test's content and the construct it is intended to measure'.[19] For an SCT score to represent a legitimate measure of clinical data interpretation (CDI) under conditions of uncertainty, the test content must, *ex hypothesi*, include problems that are *ill-defined* and *authentic*.

Fournier *et al.*[20] issued guidelines for helping SCT developers prepare test items that are ill defined (i.e. imbued with a degree of uncertainty, imprecision or incompleteness). The guidelines advocate that relevant factual knowledge should be necessary – but not sufficient – for responding to the test questions. Properly fashioned SCT questions are intended to be unanswerable using formulaic or algorithmic reasoning, or pure recall of factual information. The questions are therefore tailored to probe examinees' ability to select an appropriate alternative from among several acceptable options, rather than a single correct answer from among several factually incorrect distractors.

Success in developing suitably ill-defined SCT items can, to some extent, be verified. Questions that elicit identical responses from all experts are no different from single-correct-answer or single-best-answer MCQs, and those that obtain too broad a distribution of responses from the expert panel are considered too ambiguous.[21] By contrast, optimal SCT questions are those that produce a small range of expert responses clustered around a modal answer. High-quality questions (i.e. those with content that is most consistent with the intended construct) can therefore be easily and objectively recognised.

Table 2 Example of the script concordance test scoring system

Answer	– 2	– 1	**0**	**+ 1**	**+ 2**
Number of experts who choose this answer	0	0	1	5	4
Score	0	0	1/10	5/10	4/10
Transformed score	0	0	1/5	5/5	4/5
Credit per question	0	0	0.2	1.0	0.8

Suppose a panel of 10 experts was asked to respond to the first question in the example given in Table 1, and five experts selected response + 1, four experts selected response + 2, and one expert selected response 0. The scoring for this question would be: response 0, 0.2 points (1/5); response + 1, 1 point (5/5); response + 2, 0.8 points (4/5); responses – 1 and – 2, both 0 points. An examinee's total score for the test is the sum of the credit obtained for each of the items, divided by the total obtainable credit for the test, and multiplied by 100 to derive a percentage score

The intention behind the script concordance approach is to simulate authentic conditions of medical practice, in which courses of action or lines of thinking about specific clinical problems are seldom indisputable, even among experts.[22] Although case vignettes can never reflect the full complexity of real-patient encounters, SCT makers are instructed to generate questions from representative cases seen in daily practice.[20] In some instances, audiovisual materials, including video segments, have been used to enhance the authenticity of the test-taking experience.[2,23]

Conclusion: Published guidelines for standardising the creation of authentic, ill-defined test items serve to ensure that individual SCTs legitimately probe the method's intended construct (i.e. data interpretation in contexts of clinical uncertainty). As such, the guidelines constitute an important source of content evidence, assuming they are diligently followed during SCT development and pilot testing under non-research conditions.

Response process

The 'response process' category of validity evidence entails a search for data elucidating the relationship between an assessment's intended construct and the thought processes and response actions of its examinees.[8] Current evidence for alignment between thought and response processes and the intended construct of the SCT rests on several theoretical assumptions.

The script concordance approach is conceptually linked to a model of clinical reasoning known as the 'hypothetico-deductive' (HD) method.[12] The HD method suggests that doctors tend to generate a few hypotheses early in a clinical encounter, and subsequently orient data collection towards confirming or rejecting their initial hypotheses.[24] Patterned after this model, the SCT features three columns that correspond to the stages of hypothesis generation ('If you were thinking…'), data collection ('…and then you find…') and data interpretation ('…this hypothesis becomes…'), respectively. For each SCT question, both an initial hypothesis (column 1) and a new piece of clinical information (column 2) are provided, and therefore do not require independent generation by the examinee. What remains, ostensibly, is the stage of data interpretation, in which the examinee is presumed to make a decision regarding the fit of the new data with the given hypothesis. The script concordance method is therefore meant to probe one key signpost along an accepted theoretical pathway of clinical reasoning.

However, clinical data interpretation is not a skill that can be teased apart from the medical knowledge upon which it relies.[25] The script concordance method presumes that for each SCT question, examinees mobilise knowledge structures – 'illness scripts'[26] – from their mental databases that are relevant to the given hypothesis. Script concordance hinges on an inference that examinees with more evolved illness scripts will interpret data and make decisions that increasingly concord with those of experts given the same clinical scenarios. Indeed, SCTs used in various domains of medicine have consistently demonstrated that scores tend to increase with increasing levels of training.[4,23,27,28]

There are some empirical data to support the claim that the thought processes of SCT examinees include

a judgement of fit between new clinical data and activated scripts. In one computer-based study using the script concordance format, subjects were asked to gauge the effects (i.e. more likely, less likely, no effect) of new pieces of information on a series of diagnostic hypotheses.[29] Subjects' response times were significantly faster when they were presented with clinical information that was either typical of or incompatible with the given hypothesis than when they were presented with information that was atypical. Subjects also responded more accurately when provided with typical than with atypical information. The investigators concluded that processing time and accuracy of judgement on script concordance tasks are influenced by the degree of compatibility between new clinical information and relevant activated scripts.

Conclusion: Validity evidence in support of a clear relationship between the intended construct of the script concordance method and the thought and response processes of examinees is largely theoretical and has minimal empirical substantiation.

Internal structure

Whereas content and response process evidence is gathered to ensure that test material legitimately probes an intended construct, internal structure data provide evidence that it does so in a reproducible, or reliable, manner. The internal structure category of evidence addresses key questions related to the reliability of an assessment method.[8]

Internal structure evidence for the SCT method demonstrates dependably high measures of internal consistency, with alpha coefficient values of 0.70–0.90 across an array of medical disciplines.[1,2,6,23,27,28,30,31] The method's tendency to produce high reliability estimates is partly a function of the minimal testing time required per item, which permits the efficient collection of numerous samples of examinee performance. Script concordance tests generally contain 60–90 questions (nested in 20–25 cases for optimal reliability), and can be completed in about 1 hour.[32] They are therefore designed to diminish the problem of case-specificity that has bedevilled the interpretations of scores obtained through other methods of assessment, such as patient management problems[33] or long-case clinical examinations (CEXs),[34] that address CDI over small or single samples of items.

Another source of internal structure evidence for the script concordance method comes from data pertaining to the composition of the expert panel. Gagnon *et al.*,[35] for example, determined that a panel size of at least 10–15 members is required for acceptable (i.e. $\alpha \geq 0.70$) reliability and that up to 20 members may be necessary for high-stakes examinations. Two other studies independently discovered that whether the reference panel was composed of experts directly involved in the training of the examinees had no bearing on the relative ranking of examinee scores (although absolute scores were higher when examinee responses were compared with those of their own instructors).[36,37]

Conclusion: The SCT design has yielded remarkably robust indices of internal consistency across a spectrum of medical domains, supporting the argument that in each case a single common construct is being probed. Research concerning the ideal composition of the expert panel has yielded additional supportive evidence in this category.

Relations to other variables

To the extent that a test's score represents an underlying construct, it should correlate strongly with other indicators of the same or similar constructs, and weakly with measures of unrelated constructs.[8] Validity evidence in this category can be derived by correlating scores obtained by a method of interest with those obtained by other methods of assessment.

Two studies have investigated the correlation between SCT and MCQ test scores. Collard *et al.*[38] used a common-content blueprint to develop a fact-based true/false test and an SCT intended to probe biomedical reasoning. A positive correlation between true/false test and SCT scores was found for students at earlier (Years 3 and 4; $r = 0.53$, $p < 0.0001$), but not later (Years 5 and 6; $r = 0.07$, $p = 0.64$), stages of training. The authors concluded that 'the absence of any significant correlation in students in the later years may indicate that a relative independence of factual knowledge and clinical reasoning has developed with experience'.[38] In another study, Fournier *et al.*[31] found no significant correlation ($r^2 = 0.0164$, $p = 0.5905$) between scores on a 60-question, 'type C' (single best answer with four distractors) MCQ test and a 90-question (nested in 30 cases) SCT administered to a small cohort of residents in emergency medicine.

In a study designed to verify whether SCT scores obtained by medical students could predict 'clinical reasoning performance' as residents, Brailovsky

et al.[39] found moderate correlations between students' scores on an SCT administered at the end of clerkship and those obtained at the end of residency using two other methods for assessing reasoning in contexts of clinical uncertainty ($r = 0.451$, $p = 0.013$; $r = 0.447$, $p = 0.015$, respectively). In the same study, correlations between early SCT scores and later scores on an objective structured clinical examination (OSCE), the focus of which was to assess a somewhat different construct (reasoning during the performance of technical skills), were significantly weaker ($r = 0.352$, $p = 0.052$).

Conclusion: Studies thus far have detected relatively weak correlations between SCT scores and scores obtained on fact-based examinations, offering support to the claim that SCTs, at least to a degree, measure a different construct from tests probing pure recall of propositional knowledge. Note that the evidence here is sparse, relying on results from only a few studies that compared SCTs and single-correct-answer MCQ tests matched globally – but not on an item-by-item basis – for content. Moreover, correlations between SCT and MCQ scores in these studies were not corrected for attenuation and thus may appear falsely low. Evidence that SCT scores early in training predict later scores on tests probing similar constructs exists, but is also scant.

Consequences

This category explores evidence relating to the intended or unintended consequences of an assessment method.[8] Evidence concerning the effects of a method's scoring format, its procedure for determining score thresholds (e.g. pass/fail cut scores) and its impact on learning and teaching practices also falls under this category.[9]

The scoring format of the SCT is a version of the aggregate method that takes into account the variability of experts' responses to particular clinical situations.[40,41] It assumes that, for each question, the answer provided by the greatest number of panel members reflects optimal data interpretation under the given circumstances and that other panel members' answers reflect a difference of interpretation that is still clinically valuable and merits proportional credit. Under this paradigm, domain experts are considered to represent the reference standard for determining the degree of acceptability of different responses to SCT questions. The use of this type of scoring method in SCT has been justified and has been shown to be a key determinant of its discriminatory power.[42,43]

However, the SCT's scoring method has not gone uncontested. Bland *et al.*,[44] for example, showed that several alternative scoring methods – including single-best-answer approaches – reproduced the results obtained using the SCT's method of aggregate scoring. In general, the literature on the effects of differential weighting of item responses on validity has been tepid. For example, Sabers and White[45] reported negligible increments in reliability and validity as a result of weighted scoring. Haladyna[46] found that option weighting was labour-intensive and resulted in only slight gains in reliability and validity in a number of testing situations.

With regard to cut scores, the establishment of fair and transparent norm- or criterion-referenced methods[47] for determining success or failure on SCTs has not yet been described. Angoff, Ebel and other conventional standard-setting methods for tests with dichotomous scoring systems are not appropriate for establishing SCT cut scores. Charlin *et al.*[48] recently proposed a new statistical method for transforming and reporting scores that offers a common metric for gauging the performance of an SCT examinee relative to those of panel members. This method may, in future, be exploited to investigate standard setting and optimal pass/fail cut scores for SCTs under various testing conditions.

Several studies have explored the consequences of using the script concordance model for educational purposes during interactive workshops for health professionals.[49–51] In each study, script concordance-type questions were used to assess participants' competence in interpreting clinical data in diagnostic or management dilemmas in an area of concern. The exercise served as the basis for focused educational discussions between non-experts and experts attending the workshops. Pre- and post-workshop assessments (some were self-assessments) in each study suggested that the intervention led to improvements in participants' knowledge, clinical reasoning skills or practice habits.

Conclusion: Script concordance assessment has been, in several published instances, successfully exploited for its immediate instructional effects, whereby it helps to identify and supplement gaps in learners' knowledge structures. Little is known about the longer-term educational impact of the script concordance method on teaching and learning. Furthermore, no sufficient body of procedural evidence and outcomes data with which to defend the use of tests based on the script concordance method in high-stakes examinations currently exists and questions

remain regarding optimal methods for scoring and setting standards in SCTs.

DISCUSSION

We sought to develop a coherent validity argument for the interpretation of test scores derived through use of the script concordance method. Following an approach advocated by Messick,[18] we examined published data for five categories of validity evidence: content; response process; internal structure; relations to other variables, and consequences. We found evidence relating to content, internal structure and relations to other variables in support of the validity of SCT score interpretations, although significant evidentiary gaps remained. Conversely, evidence supporting the validity of SCT scores with respect to examinee thought and response processes and educational consequences is weaker and limited.

Limitations

A potential limitation of our exercise is that it is conventionally undertaken to evaluate the validity of inferences from scores on specific instruments developed for specific purposes. However, we have argued that results derived from whole classes (or methods) of assessment lend themselves to certain global interpretations that might be useful for helping educators decide whether or not to invest in their own versions of a test, which would then require further validity verification. Our study is also limited by the relatively small body of current SCT literature, as well as our potential bias, despite our best attempts at objectivity, as investigators who are intimately involved in SCT research and development.

Implications for education and future research

Content evidence has been bolstered by published guidelines for standardising the content and process of SCT construction. The development of suitably ill-defined test items, which describe clinical situations in which there is no single best approach, is important for lending credence to SCT score interpretations. However, the fact that not all experts agree on a single best solution to a given clinical problem does not mean that no such solution exists; more research is required to address this legitimate concern regarding SCT content validity. Careful item development and panel selection are clearly crucial for ensuring that SCT response options reflect a spectrum of acceptable practices, and that the experts reflect good clinical judgement and current clinical practice. As published work on SCTs has been carried out under research conditions, it remains to be seen how SCTs will perform when developed and implemented by non-experts. Content evidence may also be strengthened by soliciting qualitative or mixed-method data from examinees and panel members about their perceptions of the authenticity of the script concordance assessment experience.

Internal structure evidence is supported by consistently high reliability estimates from published SCTs across a spectrum of medical domains. Evidence in this category could, however, be reinforced by test–retest estimates of reliability and by generalisability studies examining the decomposition of sources of variance in SCT (e.g. errors attributable to items and item–examinee interactions versus errors attributable to answer key generation by the expert panel). With respect to the effects of panel composition on reliability, research into how expert panels that contain widely deviant responders (i.e. those with aberrantly low total scores on an SCT, or those with outlying responses to particular SCT questions) should be treated is lacking and might provide important additional evidence in this category.

Evidence from the ‘relations to other variables’ category offers some support to the hypothesis that SCTs probe a construct that diverges from that probed through most MCQ tests. Research thus far has focused on comparisons of SCTs with MCQs in which one answer is identified as clearly and unambiguously better than its alternatives. However, stronger correlations might be expected between scores on SCTs and other types of MCQs that, like SCTs, offer partial credit for answers judged reasonable but not necessarily optimal. Evidence in the ‘relations’ category might be further solidified by data extrapolated through a comparative multi-trait, multi-method research approach,[52] which would allow investigators to examine patterns of correlation between different methods of assessment (e.g. SCT versus MCQ) and the ‘traits’ (constructs) they purport to measure (e.g. reasoning in contexts of uncertainty versus knowledge fund) in a more rigorous manner.

A strategic research agenda for the SCT method should, however, focus on the two categories for which evidence is, to date, the least robust: thought and response processes, and consequences. At present, the evidence that SCT examinees’ thought and response processes align with the intended construct is based largely on theoretical argumentation. Whereas the inherent structure of its stimulus and

response format clearly precludes assessment of examinees' ability to generate medical hypotheses or collect appropriate data, the SCT's claim to probe clinical data interpretation as an isolated construct requires more empirical substantiation. Further comparative cognitive research, perhaps employing think-aloud or concept-mapping strategies, may shed light on the types of cognitive strategies examinees employ in approaching SCT questions. A fuller understanding of examinee thought and response processes is critical for helping educators to diagnose and remediate trainees who perform poorly on an SCT.

Evidence relating to consequences, or educational *impact*, is arguably the most important category of validity evidence.[53] However, at present little is known about the consequential aspect of validity of the script concordance method. For example, the script concordance method's presumed effect on learning (i.e. of steering learners away from the rote memorisation of 'textbook answers' towards deeper learning strategies) requires empirical corroboration. Furthermore, its accentuation of the role of uncertainty in clinical data interpretation, intended to simulate the conditions and complexities of real-life medical practice, may be counterintuitive to medical learners, particularly those accustomed to assessment under educational models in which 'right' answers tend to be extolled. The potential effects – positive and negative – of an assessment method rooted in uncertainty should be further explored.

The repercussions of the way that SCTs are scored, such that all panellist responses are considered to have intrinsic merit, are also open to speculation: is the SCT scoring system a tacit endorsement of the implication that 'experts never err' or an acknowledgement that practitioners often interpret data differently depending on their varying experiences (scripts) in health care? The SCT's scoring system introduces complexity into the scoring process, but may have the practical effect of reminding educators to articulate and model comfort with uncertainty when debriefing students after administering an SCT, or during other educational activities surrounding patient care. A study of the incremental value of the SCT's unique scoring system, weighed against the consequences of the complexity it entails, may therefore be warranted.

Although innovative methods for rendering SCT scores more meaningful for students may soon serve as the basis for setting standardised pass or fail scores,[48] the consequences of such decisions will undoubtedly lead to further questions. What opportunities exist for clinical educators to help remediate learners who demonstrate substandard SCT performance? How can SCT examinees who score poorly improve their CDI skills? These and other concerns about the consequences of SCT should be the primary focus of further investigation.

Finally, emerging paradigms in assessment indicate a shift in emphasis from the evaluation of individual methods or instruments to the evaluation of entire assessment programmes.[54] To date, no data exist regarding the contribution of SCTs to the delivery of a varied, competence-based assessment programme as a whole. With its emphasis on the application of knowledge, the SCT assesses trainees' competence at the 'knows how' level of Miller's pyramid.[55] As such, it has the potential to complement other assessments situated at both lower (e.g. MCQs, 'knows') and higher (e.g. OSCEs, 'shows how'; multi-source feedback, 'does') levels of Miller's pyramid. Evidence testifying to the role of the script concordance method – among a measured blend of other methods – within structured assessment programmes would further bolster the validity argument in favour of its adoption.

Contributors: SL was chiefly responsible for the study conception and design, data acquisition and interpretation, and the drafting of the manuscript. BC made substantial contributions to the study conception and to data acquisition and interpretation. DAC made substantial contributions to the study conception and design, and to data interpretation. CC made substantial contributions to the study conception, and to data acquisition and interpretation. CPMvdV contributed substantially to the conception of the study and to data interpretation. All authors contributed to the critical revision of the manuscript and approved the final version for publication.
Acknowledgements: the authors wish to thank Dr Valerie Ruhe for her insightful comments during the preparation of this manuscript.
Funding: this project was supported by a Training Award from the Fonds de la Recherche en Santé du Québec (FRSQ) to SL.
Conflicts of interest: none.
Ethical approval: not applicable.

REFERENCES

1 Charlin B, Brailovsky CA, Leduc C, Blouin D. The diagnostic script questionnaire: a new tool to assess a specific dimension of clinical competence. *Adv Health Sci Educ* 1998;**3**:51–8.
2 Brazeau-Lamontagne L, Charlin B, Gagnon R, Samson L, van der Vleuten C. Measurement of perception and

interpretation skills during radiology training: utility of the script concordance approach. *Med Teach* 2004;**26**:326–32.
3 Cohen LJ, Fitzgerald SG, Lane S, Boninger ML. Development of the seating and mobility script concordance test for spinal cord injury: obtaining content validity evidence. *Assist Technol* 2005;**17**:122–32.
4 Khonputsa P, Besinque K, Fisher D, Gong W. Use of script concordance test to assess pharmaceutical diabetic care: a pilot study in Thailand. *Med Teach* 2006;**28** (6):570–3.
5 Llorca G. Evaluation de résolution de problèmes mal définis en éthique clinique: variation des scores selon les méthodes de correction et selon les caractéristiques des jurys [Ill-defined problem assessment in clinical ethics: score variation according to scoring method and jury characteristics]. *Pédagogie Médicale* 2003;**4**:80–8.
6 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**:248–51.
7 Brownell AK. The script concordance test. *Can J Neurol Sci* 2009;**36**:272–3.
8 Cook DA, Beckman TJ. Current concepts in validity and reliability for psychometric instruments: theory and application. *Am J Med* 2006;**119**:166.e7–166.e16.
9 Downing SM. Validity: on meaningful interpretation of assessment data. *Med Educ* 2003;**37**:830–7.
10 Kane M. An argument-based approach to validity. *Psychol Bull* 1992;**112**:527–35.
11 Case SM, Swanson DB. *Constructing Written Test Questions for the Basic and Clinical Sciences*. Philadelphia, PA: National Board of Medical Examiners 1998.
12 Charlin B, van der Vleuten C. Standardised assessment of reasoning in contexts of uncertainty: the script concordance approach. *Eval Health Prof* 2004;**27** (3):304–19.
13 Kreiter CD, Bergus G. The validity of performance-based measures of clinical reasoning and alternative approaches. *Med Educ* 2009;**43**:320–5.
14 Moss P. Shifting conceptions of validity in educational measurement: implications for performance assessment. *Rev Educ Res* 1992;**62** (3):229–58.
15 Williams R, Klamen D, Hoffman R. Medical student acquisition of clinical working knowledge. *Teach Learn Med* 2008;**20** (1):5–10.
16 Montgomery K. *How Doctors Think: Clinical Judgement and the Practice of Medicine*. Oxford: Oxford University Press 2006.
17 Wainer H, Mee J. On assessing the quality of physicians' clinical judgement: the search for outcome variables. *Eval Health Prof* 2004;**27**:369–82.
18 Messick S. Validity. In: Linn RL, ed. *Educational Measurement*, 3rd edn. New York, NY: Macmillan 1989;13–103.
19 American Educational Research Association, American Psychological Association, National Council on Measurement in Education. *Standards for Educational and Psychological Testing*. Washington, DC: AERA 1999.
20 Fournier JP, Demeester A, Charlin B. Script concordance tests: guidelines for construction. *BMC Med Inform Decis Mak* 2008;**8**:18.
21 Meterissian S. A novel method of assessing clinical reasoning in surgical residents. *Surg Innov* 2006;**13**:115–9.
22 Charlin B, Boshuizen H, Custers E, Feltovitch P. Scripts and clinical reasoning. *Med Educ* 2007;**41**:1178–84.
23 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**:326–31.
24 Elstein AS, Shulman LS, Sprafka SA. *Medical Problem Solving: An Analysis of Clinical Reasoning*. Cambridge, MA: Harvard University Press 1978.
25 Boshuizen H, Schmidt H. The development of clinical reasoning expertise. In: Higgs J, Jones M, eds. *Clinical Reasoning in the Health Professions*, 2nd edn. Oxford: Butterworth Heinemann 2005;15–22.
26 Feltovich PJ, Barrows HS. Issues of generality in medical problem solving. In: Schmidt H, De Volder H, eds. *Tutorials in Problem-based Learning: A New Direction in Teaching the Health Professions*. Assen: Van Gorcum 1984;128–42.
27 Carriere B, Gagnon R, Charlin B, Downing S, Bordage G. Assessing clinical reasoning in paediatric emergency medicine: validity evidence for a script concordance test. *Ann Emerg Med* 2009;**53** (5):647–52.
28 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.
29 Gagnon R, Charlin B, Roy L, St-Martin M, Sauve E, Boshuizen HPA, van der Vleuten CPM. The cognitive validity of the script concordance test: a time processing study. *Teach Learn Med* 2006;**18** (1):22–7.
30 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* 2001;**41**:227–33.
31 Fournier JP, Thiercelin D, Pulcini C, Alunni-Perret V, Gilbert E, Minguet JM, Bertrand F. Clinical reasoning assessment in emergency medicine: script concordance tests are more efficient to detect clinical experience than rich-context multiple-choice questions. *Pédagogie Médicale* 2006;**7**:20–30.
32 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.
33 Norcini JJ, Swanson DB, Grosso LJ, Webster GD. Reliability, validity, and efficiency of multiple-choice question and patient management problem item formats in assessment of clinical competence. *Med Educ* 1985;**19**:238–47.
34 Wass V, Jones R, van der Vleuten CPM. Standardised or real patients to test clinical competence? The long case revisited. *Med Educ* 2001;**35**:321–5.

35 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**:284–91.
36 Charlin B, Gagnon R, Sauvé 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**:49–53.
37 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**:522–7.
38 Collard A, Gelaes S, Vanbelle S, Bredard S, Defraigne JO, Boniver J, Bourguignon JP. Reasoning versus knowledge retention and ascertainment throughout a problem-based learning curriculum. *Med Educ* 2009;**43**:854–65.
39 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 exploratory study on the Script Concordance Test. *Med Educ* 2001;**35**:430–6.
40 Norcini JJ, Shea JA, Day SC. The use of the aggregate scoring for a recertification examination. *Eval Health Prof* 1990;**13**:241–51.
41 Norman GR. Objective measurement of clinical performance. *Med Educ* 1985;**19**:43–7.
42 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**:150–6.
43 Charlin B, Gagnon R, Pelletier J, Coletti M, Abi-Rizk G, Nasr C, Sauve E, van der Vleuten CPM. Assessment of clinical reasoning in the context of uncertainty: the effect of variability within the reference panel. *Med Educ* 2006;**40**:848–54.
44 Bland A, Kreiter C, Gordon J. The psychometric properties of five scoring methods applied to the Script Concordance Test. *Acad Med* 2005;**80**:395–9.
45 Sabers DL, White GW. The effect of differential weighting of individual item responses on the predictive validity and reliability of an aptitude test. *J Educ Meas* 1970;**6** (2):93–6.
46 Haladyna TM. Effects of empirical option weighting on estimating domain scores and making pass/fail decisions. *Appl Measur Educ* 1990;**3** (3):231–44.
47 Norcini J, Guille R. Combining tests and setting standards. In: Norman GR, van der Vleuten CPM, Newble DI, eds. *International Handbook of Research in Medical Education*. Dordrecht: Kluwer Academic Publishers 2002;811–34.
48 Charlin B, Gagnon R, Lubarsky S, Lambert C, Meterissian S, Chalk C, Goudreau J, van der Vleuten CPM. Assessment in the context of uncertainty using the script concordance test: more meaning for scores. *Teach Learn Med* 2010;**22** (3):180–6.
49 Devlin J, Marquis F, Riker R, Robbins T, Garpestad E, Fong J, Didomenico D, Skrobik, Y. Combined didactic and scenario-based education improves the ability of intensive care unit staff to recognise delirium at the bedside. *Crit Care* 2008;**12**(1):R19 .
50 Labelle M, Beaulieu M, Paquette D, Fournier C, Bessette L, Choquette D, Rahme E, Thivierge RL. An integrated approach to improving appropriate use of anti-inflammatory medication in the treatment of osteoarthritis in Quebec (Canada): the CURATA model. *Med Teach* 2004;**26** (5):463–70.
51 Petrella R, Davis P. Improving management of musculoskeletal disorders in primary care: the Joint Adventures Program. *Clin Rheumatol* 2007;**26**: 1061–6.
52 Campbell DT, Fiske DW. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychol Bull* 1959;**56**:81–105.
53 van der Vleuten C. The assessment of professional competence: developments, research and practical implications. *Adv Health Sci Educ Theory Pract* 1996;**1**:41–67.
54 van der Vleuten C, Schuwirth LW. Assessing professional competence: from methods to programmes. *Med Educ* 2005;**39**:309–17.
55 Miller G. The assessment of clinical skills/competence/performance. *Acad Med* 1990;**65** (9):63–7.

Received 13 May 2010; editorial comments to authors 22 June 2010, 9 August 2010; accepted for publication 25 August 2010