2013, 1–10, Early Online MEDICAL TEACHER

AMEE GUIDE

Script concordance testing: From theory to practice: AMEE Guide No. 75

STUART LUBARSKY[1], VALÉRIE DORY[2], PAUL DUGGAN[3], ROBERT GAGNON[4] & BERNARD CHARLIN[4]

[1]McGill University, Canada, [2]Universite catholique de Louvain, Belgium, [3]University of Adelaide, Australia, [4]University of Montreal, Canada

Abstract

The script concordance test (SCT) is used in health professions education to assess a specific facet of clinical reasoning competence: the ability to interpret medical information under conditions of uncertainty. Grounded in established theoretical models of knowledge organization and clinical reasoning, the SCT has three key design features: (1) respondents are faced with ill-defined clinical situations and must choose between several realistic options; (2) the response format reflects the way information is processed in challenging problem-solving situations; and (3) scoring takes into account the variability of responses of experts to clinical situations. SCT scores are meant to reflect how closely respondents' ability to interpret clinical data compares with that of experienced clinicians in a given knowledge domain. A substantial body of research supports the SCT's construct validity, reliability, and feasibility across a variety of health science disciplines, and across the spectrum of health professions education from pre-clinical training to continuing professional development. In practice, its performance as an assessment tool depends on careful item development and diligent panel selection. This guide, intended as a primer for the uninitiated in SCT, will cover the basic tenets, theoretical underpinnings, and construction principles governing script concordance testing.

Introduction

The script concordance test (SCT) is used in health professions education to assess a specific aspect of clinical reasoning competence: the ability to interpret medical information under conditions of uncertainty (Charlin et al. 1998). It has demonstrated favorable psychometric qualities (construct validity, reliability, and feasibility) in research conducted across a variety of health science disciplines (Llorca 2003; Cohen et al. 2005; Sibert et al. 2006; Ramaekers et al. 2010; Deschênes et al. 2011), and across the spectrum of health professions education from undergraduate (e.g. Humbert et al. 2011) through postgraduate (e.g. Meterissian 2006) and continuing professional development (Goulet et al. 2010). Its theoretical underpinnings, rooted in script theory from cognitive psychology, are the subject of ongoing scholarly inquiry (Kreiter 2012; Lubarsky et al. 2012). Procedures for diligent construction of SCTs have been developed (Fournier et al. 2008) and systematically reviewed (Dory et al. 2012).

This guide is intended for an audience of health professions educators who have little or no familiarity with script concordance testing or its underlying rationale. Its goal is to orient the reader toward the basic tenets, theoretical concepts, and construction principles governing script concordance testing. In the first part, a general overview of the script concordance approach will be provided. In the second part, the theoretical foundation of the test format will be discussed. In the third part, practical, evidence-based recommendations for test construction will be presented.

General overview: Test principles

Design features

The SCT is a written test for assessing reasoning under conditions of uncertainty. In an SCT, examinees are presented brief clinical scenarios, followed by a series of questions soliciting judgments about diagnostic possibilities or management options when new elements of information are provided. Although sufficient clinical context is given to allow meaningful decisions to be made, a certain amount of uncertainty, imprecision, or incompleteness is deliberately embedded in each case in order to simulate the ambiguous conditions that often characterize real-life clinical encounters.

In addition to its reliance on ill-defined clinical problems, the SCT has two other key design features. The first is that the response format reflects the way medical information is often processed in challenging problem-solving situations, according to established theoretical models of knowledge organization and clinical reasoning derived from cognitive psychology and medical education research. The second is that, in contrast to most conventional forms of assessment, there are no single correct answers to SCT questions. Instead, several responses to each of the test's questions may be considered acceptable, as determined independently by members of a reference panel of experienced clinicians selected from a given discipline or knowledge domain to set the test's scoring key.

Correspondence: S. Lubarsky, McGill Centre for Medical Education, McGill University, Lady Meredith House, 1110 Pine Avenue West, Montreal, QC H3A 1A3, Canada. Tel: 1 514 3983352; fax: 1 514 3987246; email: stuart.lubarsky@mcgill.ca

ISSN 0142–159X print/ISSN 1466–187X online/13/030001–10 © 2013 Informa UK Ltd.
DOI: 10.3109/0142159X.2013.760036

Practice points

- The SCT is used to assess a specific aspect of clinical reasoning competence: the ability to interpret clinical data under conditions of uncertainty.
- The SCT has three key design features: (1) examinees are faced with ill-defined clinical situations and must choose between several realistic options; (2) the response format reflects the way information is processed in complex problem-solving situations and (3) scoring takes into account the variability of responses of experts to clinical situations.
- The SCT builds on the principles of illness script theory, which emerged from the cognitive psychology literature out of a larger debate about the nature and development of expertise.
- Evidence supporting the validity, reliability, and feasibility of SCT is derived from research conducted in many branches of the health sciences and across the spectrum of health professions education from pre-clinical training to continuing professional development.
- Practical, evidence-based recommendations exist to guide the construction of a SCT.

Stimulus and response format

The test stimulus consists of a short clinical scenario, followed by a set of questions consisting of three parts. The first part ("If you were thinking of . . . ") provides a hypothesis in the form of a diagnostic possibility, an investigative option, or a therapeutic alternative. 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 hypothesis. The question is answered in the third part (". . . this hypothesis becomes:"), which contains a Likert-type response scale (usually 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 items are shown in Figure 1.

Scoring system

SCT questions are designed to avoid having single "correct" or "consensus" answers. Instead, scoring of the SCT is based on an aggregate method that takes into account the observed variability of responses of experts to particular clinical situations (Norman 1985; Norcini et al. 1990). The SCT scoring scheme assumes that, for each question, the answer provided by the greatest number of panel members (i.e. the modal answer) may be considered "gold standard" reasoning under the given circumstances, while other panel members' answers reflect a difference of interpretation that may still be clinically valuable and worthy of partial credit. Thus, in contrast to most conventional assessment tools, the SCT employs a scoring system that acknowledges an important reality in clinical practice: that even experienced clinicians often interpret data, make judgments, and respond to uncertain clinical situations in ways that vary (within an acceptable range of medical practice) (Grant & Marsden 1988).

Theoretical underpinnings: Test rationale

What are ''Scripts''?

Script theory, rooted in cognitive psychology, proposes an explanation for how information is stored in and retrieved from the human mind to influence individuals' interpretation of objects and events in the world (Schank & Abelson 1977). Applied to the health sciences, script theory suggests that medical knowledge is organized into specialized knowledge structures called "scripts" that link relevant clinical and pathophysiological information about broad diagnostic categories (e.g. cardiovascular disease), specific illnesses (e.g. myocardial infarction), or even individual patients (e.g. the case of Mr Jones) (Charlin et al. 2000). Medical scripts, referred to as "illness scripts", begin to form during the very first clinical encounter, and become updated, restructured, tailored, pruned, and refined with experience (Schmidt et al. 1990). Mature illness scripts should be conceptualized not as expansive sets of "loosely-hanging" facts, but as richly organized networks of knowledge that permit rapid interpretation and efficient action in the face of clinical problems (Feltovitch & Barrows 1984).

Scripts and clinical reasoning

According to script theory, during each clinical encounter early signals derived from the patient and the clinical setting automatically activate a small set of pertinent illness scripts in a clinician's mind (Charlin et al. 2007). Illness scripts frame the clinician's expectations about which signs, symptoms and background characteristics the patient is likely (or not) to exhibit. The clinician's expectations are, in essence, "hypotheses" that can be evaluated through further focused data collection and interpretation. Each new piece of information gathered (e.g. historical information, physical examination findings, test results) can be interpreted as being supportive of, counter to, or having no effect on a given hypothesis. Clinical data interpretation, then, entails a clinician's evaluation of the "fit" between expected and actual clinical information toward the aim of accepting or rejecting clinical hypotheses.

The clinician's search for "fit" between expected and actual clinical features of a case can proceed in two complementary, often interactive ways: (1) through use of a "nonanalytic" type of reasoning process that relies on recognizing associative patterns, making rapid judgments, and appreciating the overall "gestalt" of the case and (2) through use of a slower, "analytic" type of reasoning process that relies on deliberate hypothesis testing and deductive thinking (Croskerry 2009). At any given moment during a clinical encounter, the cognitive strategy engaged to interpret clinical data depends in part on the quality and relevance of the clinician's activated scripts (which, in turn, depends on the clinician's prior knowledge and experience), and in part on the complexity and ambiguity of the clinical problem at hand (Mamede et al. 2007). Regardless

(a) *Judgment type: Diagnosis*
A 58-year-old woman presents to the emergency department with a two-week history of intermittent vertigo. She feels well between episodes.

If you were thinking of:	And then you find:	This diagnosis becomes:				
Q1. Benign paroxysmal positional vertigo	Episodes of vertigo last 30 minutes	–2	–1	0	+1	+2
Q2. Transient ischemic attacks	History of hypertension	–2	–1	0	+1	+2
Q3. Meniere's syndrome	Recent surgical removal of a skin lesion	–2	–1	0	+1	+2

(b) *Judgment type: Investigation*
A 33 year-old woman with polycystic ovarian syndrome and previous pregnancy-associated hypertension has been referred for evaluation of postpartum headaches, visual disturbances, and paresthesias of the arms. Her blood pressure in your office is 180/100.

If you were thinking of:	And then you find:	This investigation becomes:				
Q4. Ordering magnetic resonance venography (MRV)	The patient's headaches worsen when she lies flat	–2	–1	0	+1	+2
Q5. Ordering a 24-hour urinary protein collection	The patient underwent spontaneous vaginal delivery 4 weeks ago	–2	–1	0	+1	+2
Q6. Performing a lumbar puncture	The patient underwent caesarian section 1 week ago	–2	–1	0	+1	+2

(c) *Judgment type: Treatment*
You have been asked to see a hypertensive 74 year-old woman on hydrochlorothiazide and aspirin 80 mg daily who experienced a 15-minute episode of slurred speech and clumsiness of the left hand. Carotid dopplers demonstrate 90% stenosis of the right internal carotid artery.

If you were thinking of:	And then you find:	This treatment becomes:				
Q7. Sending her for right carotid endarterectomy	70% stenosis of the left internal carotid artery	–2	–1	0	+1	+2
Q8. Initiating statin therapy	LDL 1.97 mmol/L (normal range: 2.00–3.40 mmol/L)	–2	–1	0	+1	+2
Q9. Replacing aspirin with clopidogrel 75 mg daily	Patient has a history of peptic ulcer disease	–2	–1	0	+1	+2

Figure 1. Examples of SCT items.

of the reasoning strategy deployed, a provisional diagnosis and management plan emerges when the clinician judges that accruing data fits sufficiently well with a particular script to explain the unfolding clinical situation and/or enable appropriate action. Box 1 presents an illustrative example of script activation and processing during a clinical reasoning task.

What does SCT measure?

As shown in Figure 2, 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/hypothesis evaluation ("...this hypothesis becomes..."), respectively, of the clinical reasoning process (Lubarsky et al. 2011). The clinical scenarios and first-column hypotheses are designed to trigger the activation of relevant illness scripts from the examinee's mental database. The examinee's task is to determine the extent to which, for each question, the new piece of clinical information provided is (or is not) typical of or consistent with the features of the script mobilized by that question. Script concordance hinges on an inference that examinees with more evolved illness scripts interpret data and make judgments in uncertain situations that increasingly concord with those of experienced clinicians given the same clinical scenarios, and that performance of these skills can be captured using a Likert-type scale (Charlin et al. 1998). The observation that SCT scores consistently tend to increase with increasing level of training supports the validity of this inference (Lubarsky et al. 2011).

Box 1. Example of script activation and processing during a clinical reasoning task.

Suppose you are asked to evaluate a patient who is having headaches. When the patient enters your office, you quickly – perhaps even subconsciously – note that she is a young woman who appears to be in some discomfort. When you ask her to describe her headache, she informs you that it "affects the left side of her head" and is "very painful". These early contextual cues, both verbal ("left-sided", "painful") and nonverbal (young woman, appearance of discomfort, office setting), instantly call to mind your *migraine script:* the network of interconnected knowledge you have accumulated through prior experience and learning about the diagnosis and treatment of patients with migraines.

In this case, you anticipate that the patient will report experiencing "headaches" that are "severe" (invariant features of your migraine script), that are accompanied by "nausea" and "light sensitivity" (highly typical features), and that are "unilateral" (typical feature). Based on your experience, these features are all strongly-linked attributes of your migraine script, and you would easily recognize their tell-tale "migraine pattern." Your initial diagnostic hypothesis ("This represents a case of migraine") instantly would be supported by your discovery of these few pieces of clinical information, since they align well with your *a priori* expectations about how patients with migraines tend to present.

However, the unexpected findings of "fever" and "neck stiffness" would automatically trigger the mobilization of an alternate knowledge structure to your mind – your *meningitis script.* The clinical data at your disposal will now have to be interpreted in light of at least two competing scripts. Faced with this clinical problem, you will continue to gather and weigh information until you judge that the features of the actual case match the features of one of these scripts closely enough to enable you to proceed with appropriate investigations, treatment interventions, and counseling.

When your next patient enters the room, your previously active scripts are dismissed from working memory, and scripts that are pertinent to the new case immediately flood your mind.

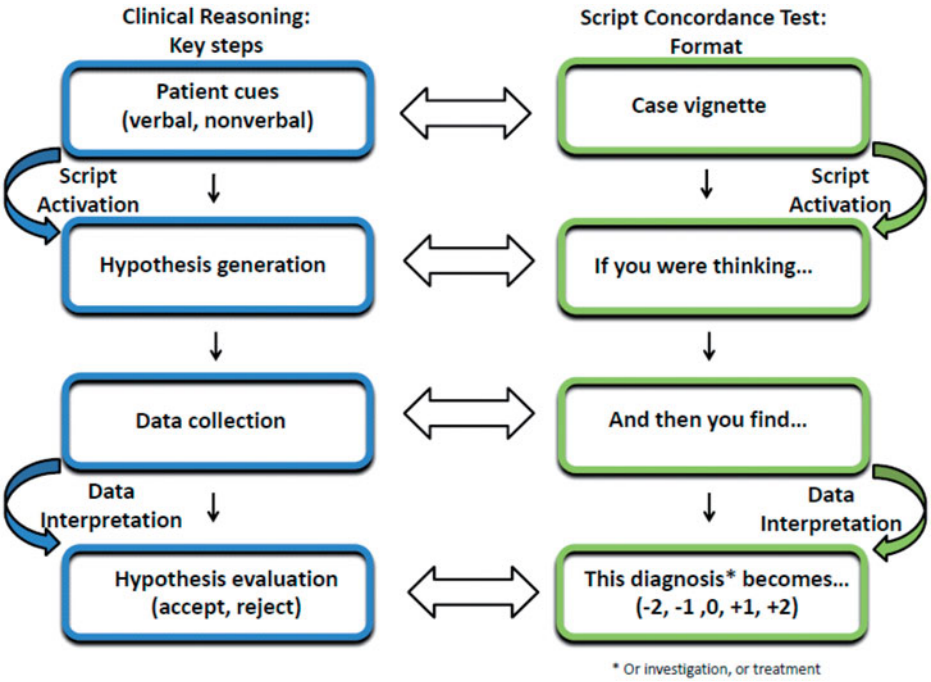

Figure 2. Relationship between key steps in the clinical reasoning process and the format of SCT items (adapted from Lubarsky et al. 2009).

What SCT does *not* measure

SCT is not designed to probe an examinee's ability to recall decontextualized, isolated facts from memory in the way that other written tests, such as multiple-choice questionnaires or short-answer essays, typically do. These types of tests are good resources for probing the breadth and depth of an examinee's knowledge base (i.e. the "amount" of knowledge possessed). The focus of SCT, on the other hand, is to explore the *structure* and *organization* of the knowledge base (i.e. how facts are *linked together* in memory and applied to context-rich, authentic clinical problems). Under this paradigm, SCT demands reasoning beyond pure retrieval of memorized knowledge: relevant factual knowledge should be necessary – but not sufficient – for responding to SCT questions.

It is, in a sense, misleading to assert that SCT is used to evaluate *clinical reasoning*; in fact, SCT concerns itself only with a specific *outcome* of the clinical reasoning process. For each SCT question, both an initial hypothesis (column 1) and a new piece of clinical information implying a data-gathering process over time (column 2) are provided. SCT, then, does not assess the examinee's ability to generate appropriate hypotheses or collect important medical information in a given clinical context. Rather, SCT focuses on the data interpretation/hypothesis evaluation stage of clinical reasoning, in which the examinee is presumed to make a decision regarding the fit of the new data with the hypothesis provided (see Figure 2). The script concordance method is therefore designed to probe one key signpost along an accepted theoretical pathway of clinical reasoning under uncertainty.

SCT in practice: Test construction

General principles

As for any assessment format, SCT construction begins with careful consideration of the intended *purpose* of the test (formative assessment, high-stakes examinations, maintenance of certification, etc.), *target group* (students, residents or interns, licensed health professionals, etc.), and *knowledge domain* (thoracic surgery, geriatrics, veterinary medicine, nursing, ethics, etc.). Subsequent test development is guided by these important concerns.

To bolster the content validity of the test results, it is advisable to create a test blueprint before developing test items (Downing & Haladyna 2006). Test blueprints are useful for ensuring that the intended knowledge domain is comprehensively covered by the test's questions. A test blueprint in general neurology, for example, might be used to ensure broad sampling of different symptom complexes (focal weakness, mental status changes, gait disturbance, etc.), specific diseases (epilepsy, multiple sclerosis, Parkinson's disease, etc.), judgment types (diagnosis, investigation, treatment, etc.), and medical settings (ambulatory care clinic, emergency department, neurological intensive care unit, etc.) across the items on the test.

Developing SCT questions

Constructing items. In an SCT, an "item" refers to a clinical scenario (called a "case," "vignette," or "case vignette") and the set of questions nested within it. For example, Figure 1 contains three SCT items, each accompanied by three questions. It is important to note (and to state explicitly in the instructions to examinees) that each question associated with a particular case should be considered *independent* of the other questions in the set. SCT items can be developed around diagnostic (Figure 1a), investigative (Figure 1b), or treatment (Figure 1c) considerations.

SCT items originate from everyday clinical experiences. Fournier et al. (2008) suggest adopting the following "key features" approach to constructing an SCT item (Figure 1a):

(1) Record a common clinical situation you have recently encountered in your clinical practice. Example: *A 58-year-old woman presents to the emergency department with a two-week history of intermittent vertigo. She feels well between episodes.* This step provides the content for the case vignette. Note that the case description should be brief, ill-defined (i.e. not all the information required to solve the problem is available), and realistic.

(2) Indicate what relevant diagnostic hypotheses or management options you would consider in this situation. Example: *benign paroxysmal positional vertigo, transient ischemic attacks, Meniere's syndrome, etc.* This step provides the content for column 1 ("If you were thinking:"), and is designed to trigger the activation of specific illness scripts in the examinee's mind. Note that the diagnostic (or investigative, or therapeutic) hypotheses must all be *plausible* (i.e. examinees should feel that the hypotheses are, indeed, reasonable considerations in the context of the given case vignette).

(3) Indicate what clinical data might help you come to an appropriate decision or course of action in this situation, and what information would have little or no effect on your reasoning. Example: *duration of vertigo, history of hypertension, accompanying auditory symptoms, recent surgical removal of a skin lesion, etc.* This step provides the content for column 2 ("And then you find:"), and simulates a data-gathering process. Note that the content devised for this column should be expected to elicit a range of positive ("+2" or "+1"), negative ("−2" or "−1"), and neutral ("0") responses on the Likert scale across the test items.

Based on our experience, we recommend that two authors assume responsibility for developing items for an SCT. The test developers should be familiar with the purpose, target audience, and content domain of the test. Prior to test administration, a preliminary draft should be sent to 2–3 independent reviewers for feedback regarding the clarity and relevance of the test items. Fournier et al. (2008) have devised a helpful survey (in English) for reviewers to consult when assessing the quality of SCT items.

Enhancing authenticity. The intent of the script concordance approach is to simulate the conditions of actual medical

Table 1. Recommended column headings and Likert-scale anchor descriptors for diagnosis, investigation, and treatment items on an SCT (adapted from Fournier et al. 2008).

Column headings Judgment type	Column 1	Column 2	Column 3		
Diagnosis	''If you were thinking of:''	''And then you find:''	''This diagnosis becomes:''		
Investigation	''If you were thinking of:''	''And then you find:''	''This investigation becomes:''		
Treatment	''If you were thinking of:''	''And then you find:''	''This treatment becomes:''		
Likert-scale anchor descriptors	−2	−1	0	+1	+2
Diagnosis	''Ruled out or almost ruled out''	''Less likely''	''Neither more nor less likely''	''More likely''	''Certain or almost certain''
Investigation (utility consideration)	''Completely or almost completely unnecessary''	''Less useful''	''Neither more nor less useful''	''More useful''	''Completely or almost completely necessary''
Investigation (risk-benefit consideration)	''Contraindicated or almost contraindicated''	''Less indicated''	''Neither more nor less indicated''	''More indicated''	''Completely or almost completely indicated''
Treatment (utility consideration)	''Completely or almost completely unnecessary''	''Less useful''	''Neither more nor less useful''	''More useful''	''Completely or almost completely necessary''
Treatment (risk-benefit consideration)	''Contraindicated or almost contraindicated''	''Less indicated''	''Neither more nor less indicated''	''More indicated''	''Completely or almost completely indicated''

practice, in which courses of action or lines of thinking about specific clinical problems are seldom indisputable, even among experts. Although the vignettes can never reflect the full complexity of real-patient encounters, SCT makers are encouraged to generate items from representative cases seen in daily practice. Audiovisual materials, including video segments, can be used to enhance the authenticity of the test-taking experience (Brazeau-Lamontagne et al. 2004; Lubarsky et al. 2009).

How many cases, how many questions?. SCTs with testing times of 60–90 min have been shown to yield adequate score reliability (Gagnon et al. 2009). Studies using classical test theory or generalizability theory have been conducted in several knowledge domains to determine the optimal number of cases and questions to include in the test. These studies indicate that, to obtain acceptable reliability (i.e. Cronbach's alpha estimates in the 0.75–0.80 range), SCTs should include approximately 25 cases with three questions nested within each case (Dory et al. 2012). The use of three questions per case in SCT is driven by theoretical as well as psychometric concerns: it has been shown that, given the finite limitations of working memory, only a small set of hypotheses is active in a clinician's mind at any given time (Kassirer 2010).

The Likert scale. Five-point Likert-type scales are commonly used in SCT, although this has been the subject of some debate (Bland et al. 2005). The anchors generally range from +2 to −2, and include a neutral point (0).[1] However, when the test is intended for use as a learning stimulus rather than as an assessment tool, it is reasonable to use a three-point Likert scale (−1, 0, +1). For example, SCT questions using three-point scales have been used as a springboard for discussion in the context of continuing health professional activities (Petrella & Davis 2007). It may also be reasonable to employ a three-point scale to evaluate novice learners, whose scripts are expected to be in the early stages of development (Kelly et al. 2012).

For each SCT item, Likert-scale anchor descriptors differ according to the type of judgment required of the examinee: diagnosis, investigation, or treatment. Table 1 suggests specific anchor descriptors for these different tasks. To encourage selection of options across the range of the Likert scale, we recommend using anchors at the extremes (+2, −2) that are not overly categorical or unequivocal (e.g. the anchor ''certain or nearly certain'' is preferable to the anchor ''absolutely certain'' or simply ''certain'').

Forming the reference panel

Panel size. Formation of a reference panel to set the test's scoring grid is a unique feature of script concordance. Gagnon et al. (2005) have shown that, for high-stakes examinations, at least 15 panel members are required to obtain adequate estimates of the reliability of scores, and only marginal benefit is gained by having more than 20 panel members. For lower-stakes examinations, fewer panel members are required, but test reliability may become compromised when panels consist of fewer than 10 members.

Panel composition. Panel composition is a key consideration during SCT development. Recall that SCT aims to compare examinees' reasoning skills with those of ''expert'' representatives of the profession or specialty examinees aspire to join. Composing an SCT reference panel is complicated by the notable lack of consensus regarding what actually constitutes ''expertise'' in a domain (Norman 2005). In the absence of standardized, evidence-based guidelines for choosing panel ''experts'', we recommend that selection decisions reflect acknowledged community standards of expertise in a given field. Criteria for SCT panel membership might, for example, include formal certification in a field, a pre-specified number

of years of practical experience in the domain of interest, an established reputation for sound clinical acumen, etc. In some instances – for example, when the test's content domain spans multiple disciplines – it is acceptable to form distinct discipline-based panels to set the response key for corresponding items on the test (Duggan & Charlin 2012).

Panel member recruitment. Recruiting 15–20 panel members to participate in an SCT is less daunting than it would seem. Anecdotally, panel members find the types of clinical problems SCT items pose appealing, as these problems tend to be representative of the common challenges they face during daily practice. Furthermore, no prior study or preparation is required to complete an SCT. Potential panel candidates are more likely to participate if full anonymity of their responses is guaranteed. Once selected, panel members are asked to complete the test independently, at their convenience, under the same time constraints imposed on examinees.

Administering the test. SCTs can be developed for administration on paper or online (Sibert et al. 2006). SCTs containing 60–90 questions (nested in 20–25 cases for optimal reliability) can be completed in about 1 h (Gagnon et al. 2009). As the test format may be unfamiliar to many examinees, SCTs should begin with clear instructions and a few practice examples (Fournier et al. 2008). Score sheets from paper-based tests can be optically read for the ease of processing.

Optimizing the test

Post hoc analysis of SCT items. One way to verify the quality of a set of SCT questions would be to pilot it on one group of respondents, evaluate its psychometric properties, and then administer an optimized version of the test to another target group of respondents. This strategy is constrained, however, by the observation that measures of test difficulty and item discrimination can vary considerably depending on the profile of respondents tested (Streiner & Norman 2008). To identify poorly performing SCT items, we therefore advocate performing an *a posteriori* item analysis of the responses of: (1) panel members and (2) examinees. This *post hoc* approach to quality assurance is a practical and justifiable method for evaluating the psychometric rigor of a test (Downing & Haladyna 2006).

Analyzing panel member responses. Within limits, response variability among the members of an SCT reference panel has been shown to be a key determinant of SCT's discriminatory power (Charlin et al. 2002, 2006). Questions that generate unanimity among panel members are no different from single correct answer multiple-choice questions (Figure 3a), and those that obtain too broad a distribution of responses may be overly ambiguous (Figure 3b). From a psychometric standpoint, ideal SCT questions are those that produce a range of expert responses clustered around a modal answer (Figure 3c). High-quality SCT questions can therefore be easily and objectively identified.

Some SCTs contain questions that elicit outlying responses from one or several of the panel members (Figure 3d). A recent study investigated the effect on test psychometrics of omitting from the test's score key responses which diverge from the modal response by more than two anchor positions[2] (Gagnon et al. 2011). For panels consisting of more than 15 members, excluding these "deviant" responses had no significant impact on the reliability of the test's results. Similarly, completely removing respondents with aberrantly low total test scores from the reference panel (i.e. those with total test scores more than two standard deviations (SD) from the panel mean) did not affect the psychometric properties of the test (Gagnon et al. 2011). Measurement error resulting from deviant panelists or deviant answers on an SCT is therefore thought to be negligible, provided the panel size is sufficiently large.

Analyzing examinee responses. Examinees' responses to the test questions can also be scrutinized to detect poorly performing items. Calculation of *item-total correlations*, an estimate of an item's discriminative capacity, is used in this step to flag problematic items. In general, items that yield negative item-total correlations, as well as items that yield item-total correlations that are positive but less than 0.05, contribute minimally or not at all to the reliability of the test scores and can be considered for removal. Poor item-total correlation, however, may simply reflect the heterogeneity of clinical competence of the panel members or of the domain tested, rather than a flaw in the item itself; in these situations, test-makers should use their best judgment to decide whether or not to discard the item.

Establishing the final version. In published studies of SCT, review of panelist and examinee responses using the strategies outlined above has led to the *post hoc* elimination of, on average, a quarter of test items (Dory et al. 2012). But test-makers beware: efforts to optimize score reliability by removing "psychometrically undesirable" questions run the risk of compromising the content validity of the test. During item review, sound judgment on the part of test-makers must be exercised in order to strike a defensible balance between these two important psychometric counterparts. To mitigate the tension between reliability and content validity, SCT developers are advised to create and administer a complement of test questions sufficiently large[3] to cover the desired content domain while still affording the leeway to eliminate certain items from the final version of the test, if necessary.

Scoring the test

Score key. Using the "optimized" version of the test, credit is assigned to each response based on how many of the experts on the panel chose 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 zero. An example of the SCT scoring system is shown in Figure 4. A scoring calculator for SCT is available online (http://www.cpass.umontreal.ca/sct.html).

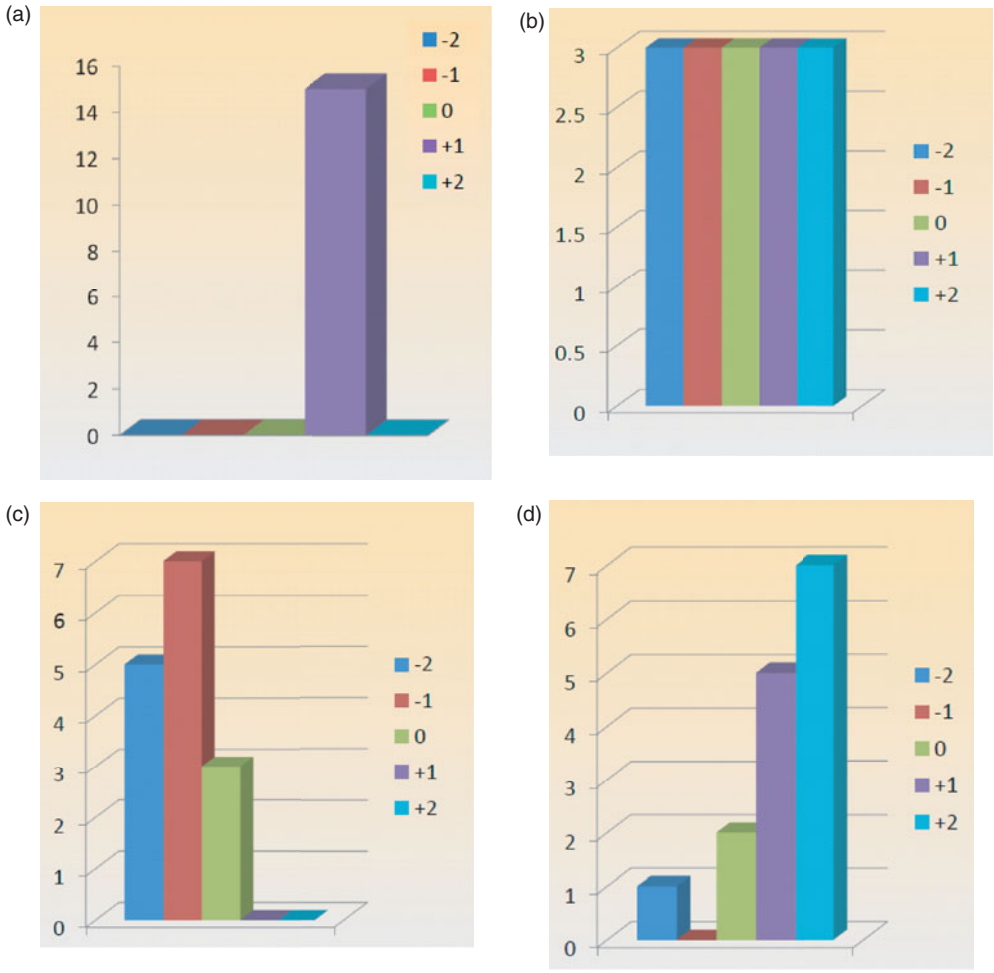

Figure 3. Verifying response variability of a single SCT question using a panel of 15 members. (a) Example of unanimity of responses within the panel. This item performs no differently than a single correct answer MCQ. (b) Example of uniform divergence of responses within the panel. This item has poor face validity and is non-discriminating. (c) Example of ideal variability of responses within the panel. Response variability is a key determinant of SCT's discriminatory power. (d) Example of a "deviant" response (in this case, the −2 response). Elimination of such responses from the answer key is not likely to affect the score's reliability.

Standard setting. For any method of assessment, setting a pass/fail cut score is an arbitrary enterprise that must take into account the purpose of the test. For SCT, it has been proposed that the pass mark should be determined in relation to the performance of experts who have sat the same test, e.g. at −2SD from the mean score of the expert reference panel (Charlin et al. 2010). For examinees at early levels of training (such as pre-clinical medical students), it may be sensible to shift the pass mark further to the left, e.g. set the pass mark at −4SD from the expert mean score (Duggan & Charlin 2012). An alternative approach is to use a different reference panel for medical students (for instance, a cohort of sixth year students may be the preferred reference panel for fifth year students).

A reasonable criticism of using the above approach for the purpose of standard setting in SCT is that the pass mark does not reflect competency in a clearly defined way. One way to resolve this issue is to calculate the borderline score using the method of Nedelsky (1954). For each question, reference panel members have selected their response from the range of 5 available Likert anchors (−2 to +2). According to the Nedelsky method, a borderline candidate would confidently exclude any option that was not chosen by any member of the reference panel, and would choose randomly from the remaining options. Further research into this and other methods of standard setting is ongoing.

Conclusion

Few standardized tools are available for use in health professions education to assess clinical reasoning competency. Existing tests, such as long-case oral exams and OSCEs, are often resource-intensive, cumbersome to administer or score, or difficult to standardize. The SCT was developed in an attempt to address this shortfall. Its aim is to probe a specific component of the clinical reasoning process: the ability to interpret clinical data, particularly under the conditions of uncertainty in which reasoning so often occurs in the clinical setting. Validity and reliability evidence pertaining to script concordance testing is mounting, and SCTs have proven to be relatively easy to create, administer, and score.

Scoring system

Suppose a panel of 15 members was asked to respond to the first question in the example given in Figure 1, and 8 members selected response +1, 5 members selected response +2, and 2 members selected response 0. The scoring for this item would be: response 0, 0.25 points (2/8); response +1, 1 point (8/8); response +2, 0.625 points (5/8); 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 questions, divided by the total obtainable credit for the test, and multiplied by 100 to derive a percentage score.

Number of panel members who chose this answer	0	0	2	8	5
Number of panel members who chose this answer divided by answer provided by the greatest number of panel members (i.e., the modal answer)	0/8	0/8	2/8	8/8	5/8
Score for this question	0	0	0.25	1	0.625

Figure 4. Scoring system.

Health professions educators may wish to consider including SCTs in their assessment programs as useful adjuncts to other traditional measures. We hope that this guide provides those wishing to do so with a rationale and a road-map.

Declaration of interest: The authors report no declarations of interest. The authors alone are responsible for the content and writing of the article.

Notes

1. As Fournier et al. (2008) point out, the zero anchor on an SCT Likert scale is not meant to be a shelter for candidates without a clear opinion, in contrast to the 0 anchor on an opinion poll that often indicates an "I don't know" response.
2. For instance, if, for a given question, the modal answer was "+1," then "−2" responses were removed from the answer key.
3. As a rule of thumb, test-makers should aim to generate one and a half (150%) times the amount of questions they plan to use in their final "optimized" version of the test (i.e. around 90–120 questions).

Notes on contributors

STUART LUBARSKY, MD, MHPE, is a member of the McGill Centre for Medical Education and an Assistant Professor of Neurology at McGill University. His research interests include clinical reasoning and assessment of competence.

VALÉRIE DORY, MD, PhD, PGDipMedEd, is a Postdoctoral Researcher at the IRSS, Université catholique de Louvain. She is funded by the Fonds de la Recherche Scientifique (FNRS). Her research interests include assessment of clinical reasoning and clinical supervision.

ROBERT GAGNON, M Psy, is the Director of the Assessment Office at the Faculty of Medicine at the University of Montréal. His research interests include assessment of clinical reasoning and test construction.

BERNARD CHARLIN, MD, PhD, is a Professor within the department of Surgery at the University of Montreal. His research field is assessment in situations of uncertainty (http://www.cme.umontreal.ca/tcs/).

References

Bland A, Kreiter C, Gordon J. 2005. The psychometric properties of five scoring methods applied to the script concordance test. Acad Med 80:395–399.

Brazeau-Lamontagne L, Charlin B, Gagnon R, Samson L, van der Vleuten C. 2004. Measurement of perception and interpretation skills along radiology training: Utility of the script concordance approach. Med Teach 26:326–332.

Charlin B, Boshuizen H, Custers E, Feltovitch P. 2007. Scripts and clinical reasoning. Med Educ 41:1178–1184.

Charlin B, Brailovsky CA, Leduc C, Blouin D. 1998. The diagnostis script questionnaire: A new tool to assess a specific dimension of clinical competence. Adv Health Sci Educ Theory Pract 3:51–58.

Charlin B, Desaulniers M, Gagnon R, Blouin D, van der Vleuten C. 2002. Comparison of an aggregate scoring method with a consensus scoring method in a measure of clinical reasoning capacity. Teach Learn Med 14:150–156.

Charlin B, Gagnon R, Lubarsky S, Lambert C, Meterissian S, Chalk C, Goudreau J, van der Vleuten C. 2010. Assessment in the context of uncertainty using the script concordance test: More meaning for scores. Teach Learn Med 22(3):180–186.

Charlin B, Gagnon R, Pelletier J, Coletti M, Abi-Rizk G, Nasr C, Sauvé E, van der Vleuten C. 2006. Assessment of clinical reasoning in the context of uncertainty: The effect of variability within the reference panel. Med Educ 40:848–854.

Charlin B, Tardif J, Boshuizen HPA. 2000. Scripts and medical diagnostic knowledge: Theory and applications for clinical reasoning instruction and research. Acad Med 75(2):182–190.

Cohen LJ, Fitzgerald SG, Lane S, Boninger ML. 2005. Development of the seating and mobility script concordance test for spinal cord injury: Obtaining content validity evidence. Assist Technol 17:122–132.

Croskerry P. 2009. A universal model of diagnostic reasoning. Acad Med 84:1022–1028.

Deschênes MF, Charlin B, Gagnon R, Goudreau J. 2011. Use of a script concordance test to assess development of clinical reasoning in nursing students. J Nurs Educ 50(7):381–387.

Dory V, Gagnon R, Charlin B. 2012. How to construct and implement script concordance tests: Insights from a systematic review. Med Educ 46(6):552–563.

Downing S, Haladyna TM. 2006. Handbook of test development. New York: Routledge.

Duggan P, Charlin B. 2012. Summative assessment of 5th year medical students' clinical reasoning by script concordance test: Requirements and challenges. BMC Med Educ 12(1):29.

Feltovich PJ, Barrows HS. 1984. Issues of generality in medical problem solving. In: Schmidt H, De Volder ML, editors. Tutorials in problem-based learning: A new direction in teaching the health professions. Assen, the Netherlands: Van Gorcum.

Fournier JP, Demeester A, Charlin B. 2008. Script concordance tests: Guidelines for construction. BMC Med Inform Decis Mak 8:18.

Gagnon R, Charlin B, Coletti M, Sauve E, van der Vleuten C. 2005. Assessment in the context of uncertainty: How many members are needed on the panel of reference of a script concordance test? Med Educ 39:284–291.

Gagnon R, Charlin B, Lambert C, Carrière B, van der Vleuten C. 2009. Script concordance testing: More cases or more questions? Adv Health Sci Educ Theory Pract 14(3):367–375.

Gagnon R, Lubarsky S, Lambert C, Charlin B. 2011. Optimization of answer keys for script concordance testing: Should we exclude deviant respondents, deviant responses, or neither? Adv Health Sci Educ Theory Pract 16(5):601–608.

Goulet F, Jacques A, Gagnon R, Charlin B, Shabah A. 2010. Poorly performing physicians: Does the script concordance test detect bad clinical reasoning? J Contin Educ Health Prof 30(3):161–166.

Grant J, Marsden P. 1988. Primary knowledge, medical education and consultant expertise. Med Educ 22:173–179.

Humbert A, Johson M, Miech E, Friedberg F, Grackin J, Seidman PA. 2011. Assessment of clinical reasoning: A script concordance test designed for pre-clinical medical students. Med Teach 33:472–477.

Kassirer J. 2010. Teaching clinical reasoning: Case-based and coached. Acad Med 85:1118–1124.

Kelly W, Durning S, Denton G. 2012. Comparing a script concordance examination to a multiple-choice examination on a core internal medicine clerkship. Teach Learn Med 24(3):187–193.

Kreiter CD. 2012. Commentary: The response process validity of a script concordance test item. Adv Health Sci Educ Theory Pract 17:7–9.

Llorca G. 2003. 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 4:80–88.

Lubarsky S, Chalk C, Kazitani D, Gagnon R, Charlin B. 2009. The script concordance test: A new tool assessing clinical judgment in neurology. Can J Neurol Sci 36:326–331.

Lubarsky S, Charlin B, Cook DA, Chalk C, van der Vleuten C. 2011. Script concordance testing: A review of published validity evidence. Med Educ 45:329–338.

Lubarsky S, Gagnon R, Charlin B. 2012. Script concordance test item response process: The argument for probability versus typicality. Adv Health Sci Educ Theory Pract 17:11–13.

Mamede S, Schmidt H, Rikers R, Penaforte J, Coelho-Filho J. 2007. Breaking down automaticity: Case ambiguity and the shift to reflective approaches in clinical reasoning. Med Educ 41:1185–1192.

Meterissian S. 2006. A novel method of assessing clinical reasoning in surgical residents. Surg Innov 13:115–119.

Nedelsky L. 1954. Absolute grading standards for objective tests. Educ Psychol Meas 14:3–19.

Norcini JJ, Shea JA, Day SC. 1990. The use of the aggregate scoring for a recertification examination. Eval Health Prof 13:241–251.

Norman GR. 1985. Objective measurement of clinical performance. Med Educ 19:43–47.

Norman GR. 2005. Research in clinical reasoning: Past history and current trends. Med Educ 39:418–427.

Petrella R, Davis P. 2007. Improving management of musculoskeletal disorders in primary care: The Joint Adventures Program. Clin Rheumatol 26:1061–1066.

Ramaekers S, Kremer W, Pilot A, van Beukelen P, van Keulen H. 2010. Assessment of competence in clinical reasoning and decision-making under uncertainty: The script concordance test method. Assess Eval High Educ 35(6):661–673.

Schank RC, Abelson R. 1977. Scripts, plans, goals, and understanding. Hillsdale, NJ: Earlbaum Assoc.

Schmidt HG, Norman GR, Boshuizen HPA. 1990. A cognitive perspective on medical expertise: Theory and implications. Acad Med 65(10):611–621.

Sibert L, Darmoni SJ, Dahamna B, Hellot MF, Weber J, Charlin B. 2006. Online clinical reasoning assessment with script concordance test in urology: Results of a French pilot study. BMC Med Educ 6:45–53.

Streiner DL, Norman GR. 2008. Health measurement scales: A practical guide to their development and use. 4th ed. Oxford: Oxford University Press.