Assessing Clinical Reasoning Skills in Scenarios of Uncertainty: Convergent Validity for a Script Concordance Test in an Emergency Medicine Clerkship and Residency

Aloysius J. Humbert, MD, Bart Besinger, MD, and Edward J. Miech, EdD

Abstract

Objectives: The Script Concordance Test (SCT) is a new method of assessing clinical reasoning in the face of uncertainty. An SCT item consists of a short clinical vignette followed by an additional piece of information and asks how this new information affects the learner's decision regarding a possible diagnosis, investigational study, or therapy. Scoring is based on the item responses of a panel of experts in the field. This study attempts to provide additional validity evidence in the realm of emergency medicine (EM).

Methods: This observational study examined the performance of medical students, EM residents, and expert emergency physicians (EPs) on an SCT in the area of general EM (SCT-EM) at one of the largest medical schools in the United States. The 59-item SCT-EM was developed for a fourth-year required clerkship in EM. The results on the SCT-EM were compared between different levels of clinical experience. Results were also compared to performance on other measures to evaluate convergent validity.

Results: The SCT-EM was given to 314 fourth-year medical students (MS4), 40 EM residents, and 13 EPs during the study period. Mean differences between the three different groups of test takers was statistically significant ($p < 0.0001$). The range of scores for the MS4s was 42% to 77% and followed a normal distribution. Among the residents, performance on the SCT-EM and the EM in-training examination were significantly correlated ($r = 0.69$, $p < 0.001$); among the MS4s who later matched into EM residency programs, performance on the SCT-EM and United States Medical Licensing Examination (USMLE) Step 2-Clinical Knowledge (Step 2-CK) exam was also significantly correlated ($r = 0.56$, $p < 0.001$).

Conclusions: The SCT-EM shows promise as an assessment that can be used to measure clinical reasoning skills in the face of uncertainty. Future research will compare performance on the SCT to other measures of clinical reasoning abilities.

ACADEMIC EMERGENCY MEDICINE 2011; 18:627–634 © 2011 by the Society for Academic Emergency Medicine

Real-world clinical decisions in emergency medicine (EM) are often made in the context of ambiguity. As medical students, interns, and residents in EM gain knowledge and experience over time, they develop greater proficiency in handling clinical situations that are ambiguous or uncertain. Such situations are generally characterized by incomplete information where there is no single clear ''correct'' answer or course of action.[1]

Assessing the development of this kind of clinical reasoning in learners of EM (as well as in medical education in general) has typically been difficult,

From the Department of Emergency Medicine, Indiana University School of Medicine (AJH, BB), Indianapolis, IN; and the Health Services Research & Development Center of Excellence on Implementing Evidence-Based Practice, Roudebush VA Medical Center (EJM), Indianapolis, IN.

Received September 7, 2010; revisions received November 28 and December 21, 2010; accepted December 25, 2010.

Presented at the Society for Academic Emergency Medicine annual meeting, New Orleans, LA, May 2009.

The authors have no relevant financial information or potential conflicts of interest to disclose.

The full SCT-EM is available upon written request from the first author and may be used without charge for educational (i.e., noncommercial) purposes.

Supervising Editor: Gary M. Gaddis, MD.

Address for correspondence and reprints: Aloysius J. Humbert, MD; e-mail: ahumbert@iupui.edu.

doi: 10.1111/j.1553-2712.2011.01084.x
ISSN 1069-6563
PII ISSN 1069-6563583

expensive, and time-consuming. The most common form of assessment in medical education, the multiple-choice question, works well when there is a single predetermined "right" answer, but is not appropriate for portraying or capturing the shades of uncertainty inherent in a clinical scenario.[2] An assessment like the Triple Jump, an essay-based examination that evaluates clinical reasoning, can be laborious and resource-intensive both to complete and to evaluate and suffers from poor evaluator inter-rater reliability.[3] Standardized patients and simulation centers provide excellent venues for assessing these kinds of clinical reasoning skills, but are generally extremely expensive to offer.[4]

As a result, medical educators have typically resorted to faculty and resident evaluations during clinical rotations to grade learners' capacities to make sound clinical decisions, including clinical reasoning under conditions of uncertainty. These evaluations tend to be subjective and dependent on many factors other than the student's ability.[5]

A relatively new assessment tool called the Script Concordance Test (SCT) presents another option. The assessment is based on the script theory of medical decision-making popularized by Schmidt et al.[6] This cognitive theory posits that medical expertise in decision-making is related to the progressive development of organized networks of knowledge called "scripts." These unique, individualized networks of experience and specialized medical knowledge expand and deepen over time, eventually reaching the point that they can be activated with little attentional effort by experts.[7,8]

The basic concept behind script theory has to do with knowledge organization. When first dealing with actual patients, medical students tend to rely on often disorganized lists of signs, symptoms, and rules memorized from their time in class, employing a basic strategy that involves including or excluding possible disorders, pathologies, etc.[8] When asked for a summary of the patient's problem, novice medical students "tend to recite endless amounts of data about clinical findings."[8] They also tend to reason causally based on their biomedical knowledge, which can be slow and laborious.[7,8]

Expert clinicians, by contrast, activate knowledge structures known as scripts. Instead of an inefficient strategy that relies on lists and rules, experts draw upon integrated *networks* of knowledge and experience related to particular illnesses and conditions. These networks work by association (instead of causality) where experts see associations and patterns that link the patient before them with rich networks of prior knowledge held together by abstract relationships. The expert clinician proceeds by association, seeking to find a script that fits the patient well, instead of trying to develop a causal theory explaining the relationship between all the patient's signs and symptoms, as a novice clinician might attempt to do. These "illness scripts" typically come with specific hypotheses about the patient's condition and a set of "attributes" (or "slots") that have "default values" defining what is "expected" and what is "unusual" for that particular hypothesis. As part of this process, clinicians seek out additional information to help verify which script is most consistent with the clinical presentation before them.[8] Experts proceed differently in terms of the order in which they explore the various "attributes," but generally end up with similar conclusions.[7,8]

In a clinically ambiguous situation, two or more scripts are activated. As the expert clinician continues to gather and process information about attributes, certain scripts gain additional weight, others might attenuate or disappear, and yet more scripts might be activated. The expert works through and processes multiple scripts, often juggling them simultaneously in an effort to consider multiple possibilities or hypotheses at once. Clinicians constantly revise their scripts in light of new information, making qualitative judgments about the conditional probability of a particular hypothesis as new data continue to surface.

This process of clinical reasoning can be considered Bayesian in the way it deals with conditional probabilities. In a 2005 article entitled "Why clinicians are natural Bayesians," Gill et al. described how clinicians apply Bayesian reasoning in developing and revising differential diagnoses: "Clinicians don't calculate a running tally of likelihood ratios as they evaluate patients. Rather, they interpret each positive result as 'somewhat more suggestive' of the disease and each negative test as 'somewhat less suggestive' and conceptualize the pre-test and post-test odds in qualitative rather than quantitative terms."[9] In a similar vein, the format of the SCT assesses underlying processes of clinical reasoning by asking test takers to state how a new piece of clinical information in a specific scenario would affect their confidence level vis-à-vis a prior working hypothesis.

The SCT format was developed by Bernard Charlin.[10] SCT items start with a short clinical vignette that is followed by a series of proposed diagnoses, investigational studies, or therapeutic interventions that a clinician might consider in those circumstances. The learner is then given one additional piece of information to the case and asked what the effect of that information would be on his or her clinical reasoning related to the proposed diagnosis, test, or therapy. This is a cognitive task that involves making qualitative judgments based on conditional probabilities. Test takers indicate their qualitative judgments for each item using a five-point Likert scale that ranges from −2 to +2. The text descriptions for the anchors vary depending on the type of question being asked.

The SCT is primarily concerned with the underlying processes by which test takers make decisions.[7,8] As a result, the particular selection of an individual test taker on an individual SCT item is less important than the overall basket of selections made by the test taker, which collectively provides a measure of how closely his or her decision-making processes align with those of expert clinicians. The answers test takers give on an SCT test are not ends in themselves, as might be the case with a traditional multiple-choice test, or with questions about evidence-based medicine, where the objective is usually to determine the single correct outcome. Instead, responses on an SCT act as proxies, or "windows," on the knowledge networks of test takers, and how effectively and efficiently these scripts can be deployed when faced with ambiguous clinical problems.

Medical students do not start to develop scripts until they start seeing actual patients.[7,8] The SCT thus assesses how similar the test takers ''organization of knowledge'' is to those of expert clinicians responding to the exact same questions based on real-world scenarios. As students gain experience and specialized medical knowledge, they develop more organized scripts that help them solve clinical problems when facing individual patients, and they also develop a greater capacity to handle multiple scripts at the same time.

Given its structure, the SCT is heavily dependent on the participation of bona fide expert clinicians when validating a particular version of a test. The scoring matrix (i.e., answer key) for an SCT is developed by giving the test to a panel of at least 10 expert physicians in the content area for the examination[11] who have good overall clinical experience in the field being tested.[12]

Experts take the SCT independently of one another and send their results back to the test developers, who compile the responses. While it is theoretically possible for an expert or group of experts to be ''wrong'' when answering a particular SCT item—perhaps a particularly gifted medical student understands a specific clinical situation better than the experts—the relatively large size of the expert panel and the relatively large number of SCT items (often more than 50 questions) help mitigate the effect of any potential aberrations that might arise.

Furthermore, because expert clinicians complete the SCT independently of one another and without discussion, the ''aggregate'' method supports diversity in possible responses and avoids the emergence of a dominant ''groupthink'' paradigm. This potential issue was examined in a study by Charlin et al. where responses to SCT items given by expert clinicians working independently were compared with responses to SCT items using a more traditional expert consensus model. The study found that 59% of answers given independently by the experts in the aggregate model differed from the answers given by the experts when group consensus was achieved.[13]

Another consideration when developing the expert panel is the overall performance of the expert clinicians on the SCT itself. The aggregate method for the SCT typically involves asking a disparate set of expert clinicians to voluntarily complete an SCT. When the tests are returned, it sometimes occurs that an individual clinician responded very differently overall to the SCT than the other expert clinicians; when the responses are scored using the other experts' responses to develop an answer key, an individual clinician may score considerably lower than his or her colleagues. In such cases, there are several possible explanations: a busy clinician may not have fully understood the nontraditional format of the SCT, may not have taken the SCT very seriously, may not have been feeling well when completing the SCT, or perhaps had a form of expertise that did not translate well into the SCT format. In any event, individual clinicians can be omitted from the expert panel on the basis of SCT performance if they score more than two standard deviations (SDs) below the expert mean.[12]

The SCT scoring matrix is based on the distribution of expert responses for each item. Full credit is awarded for the modal answer (i.e., the most commonly selected answer); partial credit for alternative responses is awarded based on the relative number of experts who selected that particular answer. Credit is thus awarded based on the degree of concordance with the expert panel's responses. More detailed and specific descriptions and guidance on writing an SCT have been previously published.[12]

The SCT has been studied in various disciplines looking at performance of both residents and medical students. There is a clear improvement in performance as a learner gains clinical experience.[14–17] The expert physician performs better than the resident physician, who in turn performs better than the third-year medical student. Recently, Carriere et al.[18] published a study looking at the SCT given to residents from various disciplines taking a pediatric EM rotation. They showed a significantly greater performance score for senior residents, compared to the junior residents.

The purpose of this study was to determine if the SCT focused on general EM would be able to reliably differentiate between medical students, residents, and experts in EM. The study also examined convergent validity by comparing performance on the SCT to established measures such as United States Medical Licensing Examination (USMLE) Step 2-Clinical Knowledge (Step 2-CK) exam and the American Board of Emergency Medicine (ABEM) in-training exam.

METHODS

Study Design

This was an observational study that compared the performance of fourth-year medical students (MS4) taking an EM clerkship, EM residents, and emergency physicians (EPs) on the SCT-EM. The study protocol was approved by the institutional review board of the Indiana University School of Medicine (IUSM), which ruled the study to be exempt from informed consent requirements.

Study Setting and Population

Students who took the SCT as part of the EM clerkship between February 2008 and February 2009 were included in the analysis. The majority of students were from IUSM, but students taking a visiting rotation to IUSM were also included in the analysis. The residents were all training as part of the IUSM EM residency program in February 2009. Students at IUSM are required to take and pass Step 2-CK during their fourth year of medical school.

Residents were recruited to voluntarily take the test after one of their conference days in February 2009. Forty of a total of 61 residents in the program (65%) took the examination. Residents took the required in-training examination later that same month. Students and residents were provided with identical written information regarding the test, a similar testing format, and a similar time frame to take the exam.

Study Protocol

The SCT-EM was developed by a team of two test writers (AJH and BB) who wrote 12 case vignettes along with 59 questions in the SCT format without any external financial support. The questions were categorized as diagnostic questions, investigational questions, and therapeutic questions. Likert-scale anchors were adapted from previously published papers on the SCT.[10,12] The topics for the questions were mapped to the clinical content from the IUSM EM clerkship curriculum. The development of test items started with commonly encountered clinical scenarios in EM and continued with a determination of data that would be sought to make decisions in that situation.

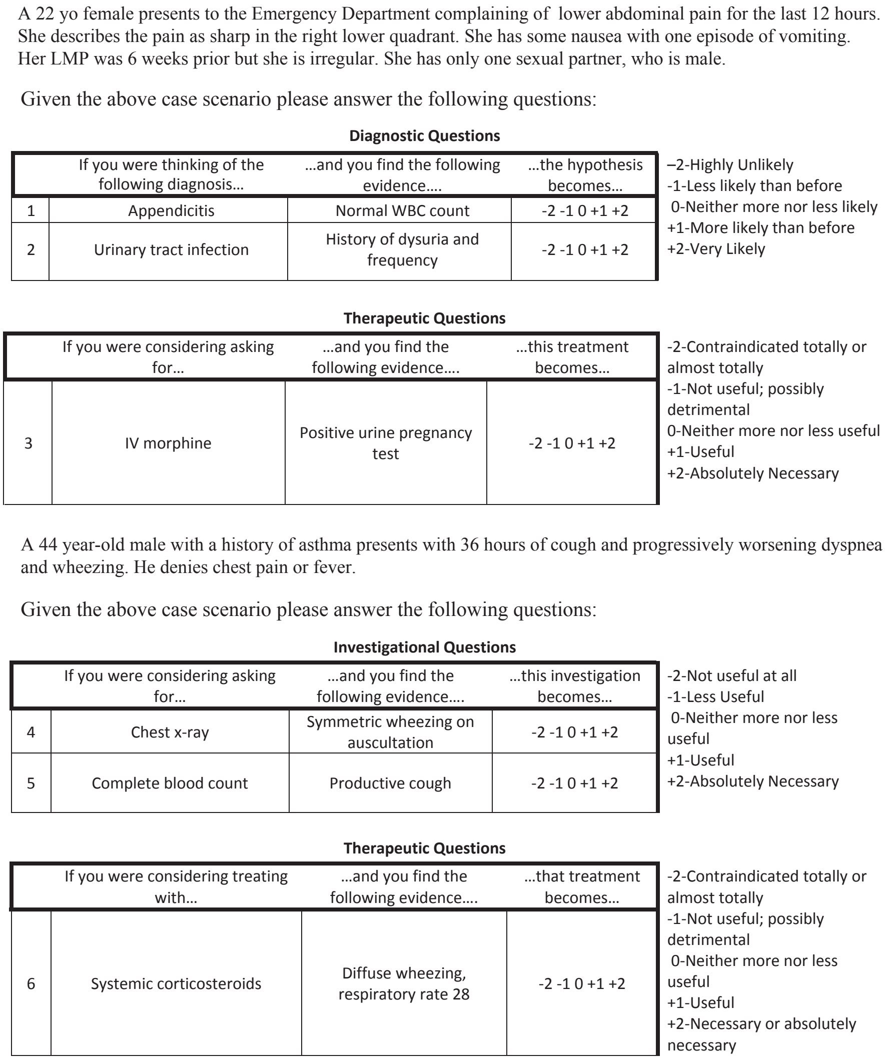
An example could be a hypothetical patient with chest pain. Data from the history and physical exam would need to be factored into the clinical decision-making process. In addition, the electrocardiogram would be a commonly used bedside test. SCT-EM items constructed around this scenario of chest pain would match a proposed diagnosis, investigational study, or therapy, with a data element that was newly obtained but may or may not be useful in making a clinical decision.

Learners respond to each SCT-EM item using a five-point Likert scale (–2, –1, 0, +1, +2) to indicate the effect of the new information on the clinical decision before them (see six sample items in Figure 1). Test items were initially developed by one author then reviewed by the other author to assess the face validity of the scenario and the test items.

The SCT-EM scoring matrix (i.e., answer key) was derived by giving the examination to a group of EM faculty for the EM clerkship. We recruited volunteers from our faculty to serve as the expert panel. We had 13 faculty agree to answer the 59 questions and submit their responses. The faculty on the SCT-EM expert panel were all EM board-certified and were from both our academic sites as well as our community hospital emergency departments where some of our students do their clinical rotations.

Measures

Responses were collated using an Excel spreadsheet (Microsoft Corp., Redmond WA). To score each



A 22 yo female presents to the Emergency Department complaining of lower abdominal pain for the last 12 hours. She describes the pain as sharp in the right lower quadrant. She has some nausea with one episode of vomiting. Her LMP was 6 weeks prior but she is irregular. She has only one sexual partner, who is male.

Given the above case scenario please answer the following questions:

Diagnostic Questions

	If you were thinking of the following diagnosis…	…and you find the following evidence….	…the hypothesis becomes…
1	Appendicitis	Normal WBC count	-2 -1 0 +1 +2
2	Urinary tract infection	History of dysuria and frequency	-2 -1 0 +1 +2

–2-Highly Unlikely
-1-Less likely than before
0-Neither more nor less likely
+1-More likely than before
+2-Very Likely

Therapeutic Questions

	If you were considering asking for…	…and you find the following evidence….	…this treatment becomes…
3	IV morphine	Positive urine pregnancy test	-2 -1 0 +1 +2

-2-Contraindicated totally or almost totally
-1-Not useful; possibly detrimental
0-Neither more nor less useful
+1-Useful
+2-Absolutely Necessary

A 44 year-old male with a history of asthma presents with 36 hours of cough and progressively worsening dyspnea and wheezing. He denies chest pain or fever.

Given the above case scenario please answer the following questions:

Investigational Questions

	If you were considering asking for…	…and you find the following evidence….	…this investigation becomes…
4	Chest x-ray	Symmetric wheezing on auscultation	-2 -1 0 +1 +2
5	Complete blood count	Productive cough	-2 -1 0 +1 +2

-2-Not useful at all
-1-Less Useful
0-Neither more nor less useful
+1-Useful
+2-Absolutely Necessary

Therapeutic Questions

	If you were considering treating with…	…and you find the following evidence….	…that treatment becomes…
6	Systemic corticosteroids	Diffuse wheezing, respiratory rate 28	-2 -1 0 +1 +2

-2-Contraindicated totally or almost totally
-1-Not useful; possibly detrimental
0-Neither more nor less useful
+1-Useful
+2-Necessary or absolutely necessary

Figure 1. Sample SCT test items from the IUSM fourth-year clerkship examination. IUSM = Indiana University School of Medicine; SCT = Script Concordance Test.

expert's SCT exam, that individual's score was removed from the distribution of expert responses and the individual test was scored using the remainder of the panel as the answer key. Prior to finalizing the answer key, one expert was removed who scored two SDs below the mean.[19]

The scoring matrix was derived by awarding full credit (one point) to the modal answer given by the expert panel. Partial credit was calculated by dividing the number of experts giving an answer by the number giving the modal answer. For example for item 1 in Figure 1, 10 of 12 experts answered ''0'' and two answered ''−1.'' Hence a full point (10/10) would be awarded for selecting ''0'' for that item and 0.2 (2/10) would be awarded for selecting ''−1''; no points would be awarded for answering −2, +1, or +2. The derivation and implementation of the scoring matrix for the six sample items is illustrated in Table 1.

Outcome measures that were analyzed included SCT-EM scores, Step 2-CK scores (for MS4), and ABEM in-training exam scores (for EM residents). Data from Step 2-CK were provided by IUSM and assigned to the appropriate student by test identification number. Data from the in-training exam were provided by the EM residency program and linked in a similar fashion. Students who later matched in EM were identified comparing data from the National Resident Matching Program match with the student roster to identify students that entered an EM program; the test identification numbers of those students were then analyzed in a separate cohort.

Data Analysis

All data analyses were performed using Minitab version 15.1.1.0 (Minitab, Inc., State College, PA), including descriptive statistics, Pearson correlation coefficients, Cronbach's coefficient alphas, two-sample unpaired t-tests, and one-way analysis of variance (ANOVA) followed by the Tukey-Kramer post hoc multiple comparison procedure. The alpha level was set at 0.05.

RESULTS

Discriminating Between EPs, Residents, and MS4s

There were 314 SCT-EM exams taken by MS4 students, 40 by EM residents, and 12 by EPs. Descriptive statistics are reported in Table 2. Mean differences between the three different groups of test-takers were statistically significant (F(2,363) = 126.0; p < 0.0001). Post hoc analysis indicated significant differences between MS4 students (mean = 60.0 [SD ± 6.2], 95% confidence interval [CI] = 59.3 to 60.7) and EM residents (mean = 70.0 [SD ± 5.4], 95% CI = 68.3 to 71.7), between MS4 students and EM faculty (mean = 79.0 [SD ± 2.9], 95% CI = 77.5 to 80.5), and between EM residents and EM faculty. These results indicated that the SCT-EM reliably discriminated among different groups of test takers with different amounts of clinical experience (i.e., experts vs. residents vs. MS4s).

Figure 2 shows the separation between the mean scores of MS4s, residents, and experts. Based on the expert SD, the resident mean score was more than

Table 1
The Derivation and Implementation of the SCT Scoring Matrix Using the Reference Panel

	Scoring Derivation Example											
	Question #1		Question #2		Question #3		Question #4		Question #5		Question #6	
Answer	# exp	CA	# exp	CA	# exp	CA	# exp	CA	# exp	CA	# exp	CA
−2	0	0	0	0	0	0	1	0.11	1	0.1	0	0
−1	2	0.2	0	0	0	0	9	1	1	0.1	0	0
0	10	1	1	0.09	7	1	2	0.22	10	1	0	0
+1	0	0	11	1	5	0.71	0	0	0	0	1	0.09
+2	0	0	0	0	0	0	0	0	0	0	11	1

# exp = number of experts giving a particular response to a question
CA=amount of credit awarded to a student giving that particular answer

	Sample Student Quiz*									
	Student #1		Student #2		Student #3		Student #4		Student #5	
Question	Answer	Points	Answer	Points	Answer	Points	Answer	Points	Answer	Points
1	−1	0.2	0	+1	0	+1	0	+1	−1	0.2
2	+1	+1	+2	0	+2	0	+2	0	+2	0
3	0	+1	+1	0.71	0	+1	0	1	+1	0.71
4	−1	+1	−1	+1	0	0.22	0	0.22	+1	0
5	+1	0	+1	0.1	0	+1	0	+1	0	+1
6	+1	0.09	+1	0.09	+1	0.09	+2	+1	+2	+1
Total pts/6	3.3/6		2.9/6		3.3/6		4.2/6		2.9/6	
Score	55%		48%		55%		70%		48%	

pts = points; SCT = Script Concordance Test.
*Scoring for five hypothetical students.

Table 2
Statistics for the SCT-EM

Group Identification	Mean (%)	SD	Sample Size	Range (%)
Expert reference panel	79.0	2.9	12	75.2–84.4
Residents	70.0	5.4	40	57.1–81.7
Fourth-year students	60.0	6.2	314	42.4–76.6
Random answer score	30.1	4.8	133	18.1–45.9

The range column shows the range of scores for each group of test takers. The range shown for the expert panel was derived by comparison of each expert's responses with the mean derived from results for pooled results from all remaining experts. Values shown for the group identified as random answer score were generated to determine the SCT-EM score that would be associated with chance.
Mean differences between the three different groups of test-takers were statistically significant: F(2363) = 126; p < 0.0001.
SCT = Script Concordance Test.

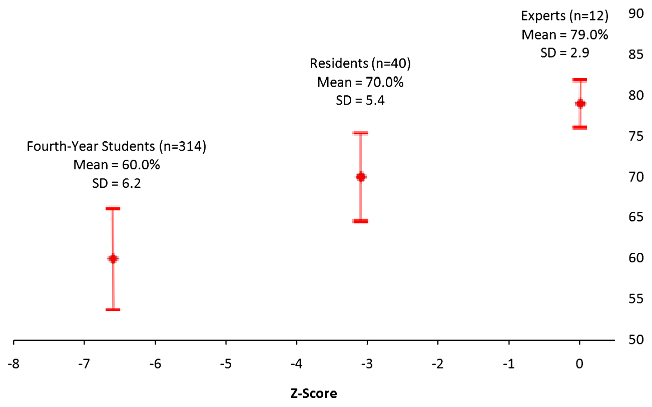


Figure 2. Mean score, SD, sample size, and z-score for three groups of SCT-EM test takers: experts, residents, and fourth-year medical students. SCT = Script Concordance Test.

three SDs below the expert mean (z-score = –3.1); the MS4 mean score was more than six SDs below the expert mean (z-score = –6.6).

Furthermore, the SCT in EM discriminated among test takers *within* the cohort of MS4 students. MS4s who successfully matched in EM residency programs (n = 15) scored significantly higher than the rest of the MS4 cohort (64.8% vs. 60.0%, t(312) = 2.96, $p < 0.004$). The mean score of 64.8% for MS4s who matched in EM residencies placed them approximately halfway between the mean scores for all MS4s (60.0%) and residents (70.0%).

The random score of 30% provided an estimate of a minimum SCT score obtained by guessing. The finding that the random mean score was 30% instead of 20%, as might be expected with test entirely consisting of items with five possible choices (i.e., the five-point Likert scale from −2 to +2), reflected the availability of answer choices that awarded partial credit. The mean random score was five SDs below the mean MS4 score.

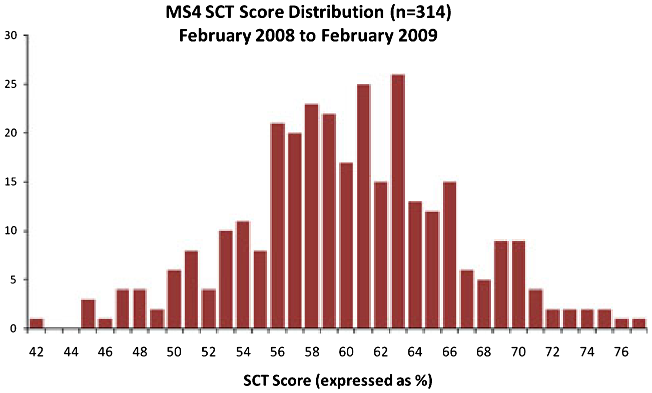


Figure 3. Distribution of MS4 SCT-EM Scores from February 2008 to February 2009 (n = 314). MS4 = fourth-year medical student; SCT = Script Concordance Test.

MS4 SCT Score Distribution

The distribution of MS4 SCT scores (n = 314) is displayed in Figure 3. The shape of the overall distribution was normal, with a smooth, continuous distribution of scores from 45% to 77%. The mean and median MS4 scores were identical at 60.0%.

A separate analysis using one-way ANOVA was conducted to evaluate whether there was any relationship between the calendar month and MS4 performance on the SCT-EM. This was completed to see if medical students who took the SCT-EM near the end of the academic year might be at an advantage relative to MS4 students who took the SCT-EM early in the academic year, as they had a few extra months of experience in other clerkships. The month-by-month analysis of SCT-EM performance by the 11 MS4 cohorts who rotated through the EM clerkship between February 2008 and February 2009 indicated no trends for SCT-EM scores to increase over time, and mean differences across the 11 administrations were not statistically significant (F(10,303) = 0.78; p = 0.652).

Convergent Validity

To assess convergent validity, the SCT-EM scores of residents were linked with individual performance on the in-training exam. It was possible to link scores for 37 of the 40 residents, and there was a statistically significant correlation between SCT performance and in-training exam scores (r(35) = 0.69, $p < 0.001$).

To assess convergent validity at the MS4 level, the SCT-EM scores of MS4s were linked to individual performance on the Step 2-CK exam. It was possible to link scores for 268 of the 314 MS4s. Overall, there was a modest but statistically significant correlation between the SCT and Step 2-CK scores (r(266) = 0.28, $p < 0.01$). However, for the 15 MS4 students who later matched in an EM residency, the correlation between the SCT and Step 2-CK scores was numerically greater (r(13) = 0.56, $p < 0.001$).

Internal Reliability

The Cronbach's coefficient alpha for the SCT-EM was calculated for the full sample of test takers and a cohort with the most EM specialized content knowledge and experience. Based on the full sample of experts ($n = 12$), residents ($n = 40$), and MS4s ($n = 314$), the alpha was 0.60. The alpha increased to 0.78 with a sample consisting of the experts ($n = 12$), residents ($n = 40$), the IUSM MS4s who later matched into EM residency programs ($n = 15$), and the visiting MS4s who later matched into EM residencies ($n = 16$).

DISCUSSION

This study was the first to include undergraduate medical students in a study of administering an SCT in the domain of EM. The performance of test takers on the SCT-EM significantly improved across levels of clinical experience, a finding that is consistent with other prior studies on the SCT.

The SCT-EM is an efficient way to evaluate a learner's ability to use information in an ambiguous clinical context. Students generally took the 59-item test in about 30 minutes. Scoring the exam is efficient as well and could be performed almost instantaneously using an Excel spreadsheet. Although not used in this study, online administration represents another option for automated SCT scoring.

This study was the first, to our knowledge, to compare an SCT with other well-established standardized assessments (the ABEM in-training exam and the Step 2-CK exam) and find strong evidence of convergent validity. This appears to underscore the role of basic medical knowledge, in addition to clinical experience, in developing and activating scripts to guide clinical reasoning processes.

Future directions for research on the SCT include looking at comparisons between SCT results and other assessment methodologies used to assess clinical reasoning ability, such as clinical ratings, objective structured clinical examinations, and oral examinations. Alternative methods of score reporting may also provide students with a better understanding of their SCT scores.

While we have focused on the use of the SCT-EM for assessment, the SCT-EM can also be used for instructional purposes. For example, in an EM clerkship, midway through a relevant lecture, a presenter could use a slide to pose an SCT-EM item to the audience. Learners could then indicate their answer (on the −2 to +2 continuum) with a show of hands, color-coded cards, or an audience response system. After noting the initial distribution of responses, the presenter could follow with large- or small-group discussion of the different rationales underlying the different answer choices.

For example, if the question that asks the test taker to consider the effect of a normal white blood cell count on the possible diagnosis of appendicitis (Figure 1) was posed in this manner, many novice learners may think that the diagnosis is highly unlikely (−2) given this new information. When the panel's responses are revealed, and with experts answering −1 or 0, a discussion of the characteristics of the white blood cell count in appendicitis including the sensitivity, likelihood ratios, and Bayesian analysis of the situation could ensue. The presenter could then repoll the audience on that particular SCT-EM item to see if the distribution of answers had changed.

The SCT format has already been used in continuing medical education as a framework for discussion, where SCT items were given to the attendees and used to discuss differences between the responses of individuals, their peers, and the expert panel.[20]

LIMITATIONS

This study was conducted at a single medical school and residency program, and the results may not necessarily be generalizable to other institutions. Administering the SCT-EM at multiple institutions would be advantageous to further strengthen the test methodology and inform ongoing research discussions about using the SCT format at multiple sites on a regional or national scale.

This study used the students' Step 2-CK scores and residents' in-training examination scores as a comparison to assess for convergent validity. These are multiple-choice examinations that focus more on the assessment of medical knowledge than clinical decision-making. Comparison to a more direct examination of clinical decision-making ability (e.g., the ABEM oral certification exam) would provide a more ideal comparison, but was not feasible in the study setting.

The internal reliability of our assessment was less than optimal with a Cronbach's alpha of 0.60 and may have been related to the relatively low number of SCT test items and the relative difficulty of the SCT for the MS4s, who had an average score of 60% on the test.

CONCLUSIONS

Overall, the use of the Script Concordance Test for emergency medicine shows much promise as an efficient measure to assess clinical reasoning in scenarios of uncertainty. The Script Concordance Test embraces the ambiguities of real-world clinical practice with its unique format, providing a viable and practical assessment to integrate into emergency medicine education.

References

1. Epstein RM. Mindful practice. JAMA. 1999; 282: 833–9.
2. Epstein RM. Assessment in medical education. N Engl J Med. 2007; 356:387–96.
3. Nendaz MR, Tekian A. Assessment in problem-based learning medical schools: a literature review. Teach Learn Med. 1999; 11:232–43.
4. Cusimano MD, Cohen R, Tucker W, Murnaghan J, Kodama R, Reznick R. A comparative analysis of the costs of administration of an OSCE (objective structured clinical examination). Acad Med. 1994; 69:571–6.
5. Herbers J, Noel G, Cooper G, Harvey J, Pangaro L, Weaver M. How accurate are faculty evaluations of clinical competence? J Gen Intern Med. 1989; 4:202–8.
6. Schmidt HG, Norman GR, Boshuizen HP. A cognitive perspective on medical expertise: theory and implication. Acad Med. 1990; 65:611–21.

7. Charlin B, Boshuizen HP, Custers EJ, Feltovich PJ. Scripts and clinical reasoning. Med Educ. 2007; 41:1178–84.
8. Charlin B, Tardif J, Boshuizen HP. Scripts and medical diagnostic knowledge: theory and applications for clinical reasoning instruction and research. Acad Med. 2000; 75:182–90.
9. Gill CJ, Sabin L, Schmid CH. Why clinicians are natural bayesians. BMJ. 2005; 330:1080–3.
10. Charlin B, Roy L, Brailovsky C, Goulet F, van der Vleuten C. The Script Concordance test: a tool to assess the reflective clinician. Teach Learn Med. 2000; 12:189–95.
11. 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.
12. Fournier JP, Demeester A, Charlin B. Script Concordance Tests: guidelines for construction. BMC Med Inform Decis Mak. 2008; 8:e18.
13. 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.
14. 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.
15. 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.
16. Sibert L, Charlin B, Corcos J, Gagnon R, Lechevallier J, Grise P. Assessment of clinical reasoning competence in urology with the Script Concordance Test: an exploratory study across two sites from different countries. Eur Urol. 2002; 41:227–33.
17. 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.
18. Carriere B, Gagnon R, Charlin B, Downing S, Bordage G. Assessing clinical reasoning in pediatric emergency medicine: validity evidence for a Script Concordance Test. Ann Emerg Med. 2009; 53:647–52.
19. Charlin B, Gagnon R, Lubarsky S, et al. Assessment in the context of uncertainty using the Script Concordance Test: more meaning for scores. Teach Learn Med. 2010; 22:180–6.
20. Petrella RJ, Davis P. Improving management of musculoskeletal disorders in primary care: the Joint Adventures Program. Clin Rheumatol. 2007; 26: 1061–6.