Assessment of clinical reasoning: A Script Concordance test designed for pre-clinical medical students

ALOYSIUS J. HUMBERT1, MARY T. JOHNSON2, EDWARD MIECH3, FRED FRIEDBERG4, JANICE A. GRACKIN5 & PEGGY A. SEIDMAN6

1Indiana University School of Medicine, USA, 2Florida State University College of Medicine, USA, 3Roudebush Veteran Affairs Medical Center, USA, 4Department of Psychiatry, Stony Brook University Medical Center, USA, 5Nassau Community College, USA, 6Department of Anesthesiology and Pediatrics, Stony Brook University Medical Center, USA

Abstract

Background: The Script Concordance test (SCT) measures clinical reasoning in the context of uncertainty by comparing the responses of examinees and expert clinicians. It uses the level of agreement with a panel of experts to assign credit for the examinee’s answers.

Aim: This study describes the development and validation of a SCT for pre-clinical medical students.

Methods: Faculty from two US medical schools developed SCT items in the domains of anatomy, biochemistry, physiology, and histology. Scoring procedures utilized data from a panel of 30 expert physicians. Validation focused on internal reliability and the ability of the SCT to distinguish between different cohorts.

Results: The SCT was administered to an aggregate of 411 second-year and 70 fourth-year students from both schools. Internal consistency for the 75 test items was satisfactory (Cronbach’s alpha = 0.73). The SCT successfully differentiated second- from fourth-year students and both student groups from the expert panel in a one-way analysis of variance ($F_{2,508} = 120.4; p < 0.0001$). Mean scores for students from the two schools were not significantly different ($p = 0.20$).

Conclusion: This SCT successfully differentiated pre-clinical medical students from fourth-year medical students and both cohorts of medical students from expert clinicians across different institutions and geographic areas. The SCT shows promise as an easy-to-administer measure of “problem-solving” performance in competency evaluation even in the beginning years of medical education.

Introduction

The Script Concordance test (SCT) is an innovative assessment instrument originally developed by Charlin et al. (2000). In comparison to multiple-choice questions (MCQ) that reliably test the ability of students to apply known solutions to well-defined problems (Gagnon et al. 2006), the SCT measure is designed to evaluate clinical reasoning in structured but uncertain diagnostic and treatment situations. Scoring is based on comparing the responses of individual test-takers with those of a panel of expert clinicians (Charlin et al. 2000). The test is based on “script theory” as it applies to the development of medical expertise (Feltovich & Barrows 1984). This theory of abstract cognitive structures posits that clinicians mobilize unique individualized networks of organized knowledge, called scripts, to process information and progress toward solutions of clinical problems (Charlin et al. 2007). Enhanced by clinical experiences, these networks expand over time and eventually adapt to specific cognitive tasks that clinicians routinely perform (Schmidt et al. 1990).

The SCT has been studied in the fields of family medicine, surgery and radiology, among others (Feltovich & Barrows 1984; Charlin et al. 2000, 2007; Gagnon et al. 2006). The reliability and construct validity of the SCT are well documented when used to differentiate advanced medical students, residents, and expert attending physicians (Sibert et al. 2002, 2005; Charlin & van der Vleuten 2004). In addition, the SCT has been used to assess the clinical reasoning skills of medical students during their clerkship years, as well as those of residents and fellows (Charlin et al. 2000). However, we found no published reports of SCT development for medical students during pre-clinical training.

The SCT described here was designed by a committee of basic science and clinical faculty members explicitly for evaluating early medical learners. This is a new type of SCT for general topics in medicine, relevant to the evaluation and training of pre-clinical students. We will argue that expanding the pool of SCT test-takers to include pre-clinical medical students has the potential to (1) provide informative evaluations of the clinical reasoning abilities of these students in the
The SCT is an innovative and practical instrument that introduces a written case about a patient that is deliberately ambiguous. When new clinical information and possible response options are presented, the answers of examinees and expert clinicians are compared, using the level of agreement to assign credit for the examinees response. The SCT has been shown to have adequate internal reliability and to differentiate medical learners, e.g., third- to fourth-year medical student, resident, expert.

This study developed a new SCT for students in the preclinical phase of education (first- to second-year medical student). The new test distinguished this cohort of learners from fourth-year students and experienced physicians.

The SCT, as a measure of a learner’s clinical reasoning skills, could be used to guide efforts at enhancing the development of problem solving competence in medical education.

context of medical uncertainty; (2) offer students feedback from a measure that may be associated with individual growth and development; and (3) add a new dimension to program assessment that may be directly relevant to “problem-solving,” one of the evolving constructs in competency evaluation (Wass et al. 2001).

The purpose of this study was to develop and validate a new SCT for pre-clinical medical students utilizing dual-institution collaboration across different regions of the USA. We describe test construction and scoring procedures, report summary test results and discuss our findings with respect to competency evaluation.

A 34 year old woman presents to the family care clinic with 24 hours of right upper quadrant pain and fever.

<table>
<thead>
<tr>
<th>If you were thinking of the following diagnosis...</th>
<th>...and you find the following evidence...</th>
<th>...this diagnosis becomes</th>
<th>+2=almost certain</th>
<th>+1= somewhat more probable</th>
<th>0 = neither less or more probable</th>
<th>–1= somewhat less probable</th>
<th>–2= almost ruled out</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hepatitis</td>
<td>Scleral icterus</td>
<td>–2 –1 0 +1 +2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pneumonia</td>
<td>Decreased breath sounds, right lower lung</td>
<td>–2 –1 0 +1 +2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Duodenal ulcer</td>
<td>Occult blood in stool</td>
<td>–2 –1 0 +1 +2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 1. Sample SCT items. SCT items from this study developed for use in undergraduate medical education, in consultation with committees of basic science and clinical faculty from both medical schools. Students were asked to indicate the effect that a piece of evidence presented in the middle column might have on a diagnostic tendency shown in the first column. The clinical reasoning process is reflected as a probability phenomenon rather than an absolute choice.
included all of the authors, was responsible for (1) finalizing the selected items, (2) administration and grading of the SCT, and (3) providing feedback to the student test-takers. This second tier of editing by the dual-institution working group assured a balance between regional viewpoints and the unifying force of common evaluation goals.

Each test item consisted of a brief, plausible patient vignette based on a common clinical complaint. The vignette was followed by a series of test questions, each of which presented medical data possibly relevant to diagnosis of the hypothetical patient. Each test question was independent of all other questions in that vignette. Questions introduced new information that was intended to increase, decrease, or have no impact on the likelihood of a specific diagnosis or treatment.

Items were discussed during working group conference calls and edited using free, collaborative, web-based software applications (e.g., Google Docs). For all items, content validity was defined by the extent to which test items are representative and relevant to targeted medical domains as established by agreement of the joint working group. Content validity assessment and item writing were done by different individuals. The final form of the test was easy to administer and score, comparable to a multiple-choice examination of similar length.

Item scoring

Unlike a conventional multiple-choice examination, the SCT awards credit for more than one answer per question. For any given test question, the modal answer of the reference panel physicians received full credit (e.g., 1 point). Partial credit for non-modal answers was based on the number of valid experts who selected the same answer for that item (i.e., the number of experts selecting that particular answer divided by the number of experts selecting the modal answer for that question; Bland et al. 2005). This is akin to clinical situations in the real world, where more than one approach is often possible, yet one particular approach may be more commonly accepted. The amount of partial credit available varied for each SCT test question and was based on the distribution of responses from the expert panel.

Given the differences in scoring between the SCT and a standard multiple choice examination, a random guessing score for the entire SCT test was expected to exceed the usual 20% correct baseline for an examination with only one correct answer for each question (five answer choices). To approximate an average score based on random guessing for our SCT, a random number generator was used to produce ten thousand values to represent answer selections. These values in total yielded 135 completed random score tests of 75 items each. From these completed tests, an average random score was calculated (Table 1) that represents the minimum score expected from random guessing.

Reference panel

The IUSM–SBUMC project’s SCT reference panel, recruited from both institutions, consisted of 30 experienced clinicians, between 2 and 15 years post-residency where data were available. Previous SCT development studies found that a minimum of 10 experts is required for item validation, and that for high-stake examinations, 15 panel members are required in order to obtain acceptable reliability estimates (Gagnon et al. 2005; Fournier et al. 2008). Our general medicine reference panel consisted of both academic- and community-based faculty from the specialty areas of family medicine, general internal medicine and emergency medicine.

A previous study found no significant differences in answer patterns between teaching and non-teaching SCT reference panels (Charlin et al. 2007). In order to minimize the anticipated time commitment for our study, each expert was assigned half of the SCT questions (37 or 38 items) from the full 75-item test. Since SCT test vignettes resemble common medical encounters, practicing physicians were assumed to be capable of answering test questions without preparation. Answers provided by the clinical experts, and corresponding SCT scores, were coded to maintain confidentiality.

Between November 2007 and January 2008, 48 hard-copy versions of the SCT tests were mailed to the expert physicians in Indiana (32) and in New York (16). Of the 48 SCT exams mailed to our experts, 33 were completed and returned, 22 from Indiana and 11 from New York (69% response rate). Eighteen of the completed tests represented the 37-item version of the SCT test, and 15 the 38-item version. According to Fournier et al. (2008), the expert panel should be made up of physicians with good overall clinical performance in the areas that the exam tests. One method that has been utilized to accomplish this is to remove the scores of panel members falling below two standard deviations (SDs) of the mean (Charlin et al. 2010). Of the 33 expert response sets, three were removed from the reference panel because the scores fell more than two SDs below the expert mean. No experts scored more than two SDs above the mean. Each SCT item was reviewed by at least 13 clinical experts,

<table>
<thead>
<tr>
<th>Group identification</th>
<th>Mean (%)</th>
<th>SD</th>
<th>Sample size</th>
<th>Range (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expert reference panel</td>
<td>80.46</td>
<td>5.27</td>
<td>30</td>
<td>69.88–90.55</td>
</tr>
<tr>
<td>Fourth-year students, IUSM</td>
<td>69.48</td>
<td>8.06</td>
<td>52</td>
<td>45.90–80.50</td>
</tr>
<tr>
<td>Fourth-year students, SBUMC</td>
<td>67.78</td>
<td>7.73</td>
<td>18</td>
<td>48.87–81.40</td>
</tr>
<tr>
<td>Second-year students, IUSM</td>
<td>60.31</td>
<td>7.83</td>
<td>296</td>
<td>33.06–77.45</td>
</tr>
<tr>
<td>Second-year students, SBUMC</td>
<td>61.44</td>
<td>7.39</td>
<td>113</td>
<td>40.97–76.15</td>
</tr>
<tr>
<td>Random answer score</td>
<td>29.90</td>
<td>4.34</td>
<td>153</td>
<td>18.64–41.40</td>
</tr>
</tbody>
</table>

Notes: The range column shows the range of student responses for each group. The range shown for the expert panel was derived by comparison of each expert’s responses with the score derived from pooled results of all remaining experts. Values shown for the group identified as random answer score were generated to determine the probable SCT score that would be associated with chance.
consistent with literature recommendations for expert panel numbers (Gagnon et al. 2005; Fournier et al. 2008).

Institutional Review Board Approval
This study protocol was approved by the Institutional Review Boards (IRBs) of both SBUMC and IUSM.

Data collection
The new SCT was administered school-wide to 350 students from IUSM and 113 students from SBUMC. At IUSM, this included all 298 beginning second-year students from nine campuses distributed around the state and a convenience sample of 52 of 275 fourth-year students. At SBUMC, the SCT was completed by all 113 second-year students and 18 of 102 fourth-year students. To our knowledge, no students at IUSM or SBUMC had taken a SCT prior to taking this exam.

The large difference in recruitment rates between second- and fourth-year students is primarily attributable to the predictable class location of the second-year students as compared to the elective rotation locations of the fourth-year students. The pencil-and-paper SCT was administered during 1-h time periods in proctored settings. Students were not expected to study or prepare for the test. Test administrators at each site received a complete packet with all needed materials, including test booklets, glossaries, and proctoring instructions.

Students took the test anonymously and scores remained confidential by coding with unique test identification numbers. Students could use these coded identification numbers to learn their scores after taking the test. To adapt the test to pre-clinical medical students, given that the focus of assessment was clinical reasoning rather than medical content knowledge, each examinee received a hard-copy reference glossary defining all medical terms used in the SCT. The student’s score report featured individual results along with means and ranges for the reference panel and student cohorts for comparison.

Data analysis
Reliability was calculated with Cronbach’s coefficient alpha (75 test items; \( n=481 \) students). Data were analyzed with descriptive statistics and one-way analysis of variance followed by Tukey post hoc tests using SPSS, version 16 (SPSS, Inc., Chicago, IL). The alpha level was set at 0.05.

Results
Of the 411 second-year and 70 fourth-year students who took the new SCT, 100% of the tests were complete and scorable. Table 1 presents descriptive statistics for the tested samples and for scores based on random answers. The internal consistency of the test was acceptable (Cronbach’s alpha = 0.73). Statistically significant differences were found between groups (reference panel physicians, fourth-year students and second-year students; \( F_{2,508}=120.4; \ p<0.0001 \)). Post hoc analysis indicated that significant differences were present between second- and fourth-year students, between second-year students and experts, and between fourth-year students and experts (\( p<0.001 \); Figure 2). No significant difference was found between Indiana and New York with respect to the mean scores of the second-year students (\( p=0.2 \)). Furthermore, the mean scores of the experts from the two states were not significantly different (\( p=0.24 \)).

Discussion
This study described the development, initial validation, and school-wide implementation at two different sites of a new SCT for pre-clinical medical students that focuses on clinical reasoning in general medicine. The SCT developed from this dual-institution collaboration demonstrated satisfactory internal consistency and significant differentiation of scores between (1) second- and fourth-year medical students and (2) expert clinicians and students. The administration of the SCT to second-year medical students also showed no significant differences in the mean scores between test-takers at IUSM and SBUMC. This latter finding supports the generalizability of the assessment instrument. As far as we are aware, this is the first SCT used to evaluate clinical problem solving for medical students in their pre-clinical years.

Differences between SCT and conventional tests
The evaluation of problem solving competency in early medical learners is a considerable challenge. Traditional multiple choice questions (MCQs) and the structured triple jump exercise used in conjunction with problem-based learning are two commonly used assessments (Nendaz & Tekian 1999; Epstein 2007). The literature identifies difficulties specific to the evaluation of problem solving ability in relation to such methods. Although a MCQ examination composed of context-rich questions may show a reasonable level of validity for evaluating complex clinical decision-making, the MCQ examination used in both basic and clinical sciences is not a substitute for direct evaluation of clinical problem solving.
measurements of clinical problem solving that might have
not attempt to compare SCT results with other ratings or
Limitations
Barrows 1984).
useful information, and generate a solution (Feltovich &
defined in part as the ability to confront questions, gather
related to the clinical competency of problem-solving,
stages of medical education. SCT performance appears to be
suggest that clinical reasoning can be assessed at the early
medical students could provide them with an indication
their progress in assessing common clinical situations.
Because SCT scores are normed on the performance of
a panel of expert clinicians, yearly SCT performance data for
medical students could provide them with an indication
of their progress in assessing common clinical situations.
Apart from medical knowledge acquisition, our findings
suggest that clinical reasoning can be assessed at the early
stages of medical education. SCT performance appears to be
related to the clinical competency of problem-solving,
defined in part as the ability to confront questions, gather
useful information, and generate a solution (Feltovich &
Barrows 1984).

SCT and direct observations of problem solving competence
Problem solving comprises a crucial skill set for practicing
physicians. For beginning medical learners, the large body of
knowledge necessary to be learned in combination with rapid
advances in clinical management suggest the utility of incor-
porating assessments that present ambiguity and change with
respect to real-life clinical problems (Maudsley & Strivens
2000). Problem-solving skill is an organized activity that can
become routine when shaped by adequate feedback (Werb &
Matear 2004). During clerkships, growth in student skill is
observed and rated subjectively using global clinical scales.
However, this type of practical feedback is dependent
upon individual observer critique, which is not available in
the pre-clinical setting.

Because SCT scores are normed on the performance of
a panel of expert clinicians, yearly SCT performance data for
medical students could provide them with an indication
of their progress in assessing common clinical situations.
Apart from medical knowledge acquisition, our findings
suggest that clinical reasoning can be assessed at the early
stages of medical education. SCT performance appears to be
related to the clinical competency of problem-solving,
defined in part as the ability to confront questions, gather
useful information, and generate a solution (Feltovich &
Barrows 1984).

Limitations
Several limitations of this study require comment. The SCT
presumably tests problem-solving ability. However, we did
not attempt to compare SCT results with other ratings or
measurements of clinical problem solving that might have
provided convergent or concurrent validity. In addition, the
analysis comparing fourth-year students at IUSM and SBUMC
was under-powered (19.8%) to detect a difference in the
two means with a one-tail test at an alpha level of 0.05. Finally,
the SCT is limited in its ability to assess how students address
psychological, interpersonal, and emotional issues when
dealing with clinical situations that involve ambiguity and
uncertainty.

Application of SCT results
The reported SCT findings may have specific applications for
medical education. Variations in testing outcomes could be
used to identify students who have difficulty with clinical
reasoning early in their training. Early identification and
remediation of these students may allow for a focused
and timely intervention to foster improvement in clinical
reasoning. Rather than using the SCT outcomes as a criterion
for advancement, or a relative ranking instrument, the
intended purpose is to provide students with personal skills
development informed by individualized feedback about
clinical reasoning ability. In addition, systematic variations in
student scores between sites or programs could also prompt a
closer look at the factors contributing to scores substantially
above or below overall averages. This process would provide
valuable feedback to instructors and program directors that
might help to inform curricular development.

A validated SCT for early medical learners may also serve as
a starting point to investigate methods that potentially facilitate
student development of problem solving skills. Although the
differential knowledge levels of the groups tested in this
study may plausibly explain a portion of the observed variance
in test scores, the finding that at least some novices performed
nearly as well as experts suggests that advanced problem-
solving skills may be involved as well. This suggests an
opportunity to ascertain factors that may be associated with
advanced clinical reasoning ability. Such information may
inform future iterations of the pre-clinical SCT and identify
strategies to guide medical students in their developing
problem solving competence.

Conclusion
Undergraduate medical assessments may be enhanced with
structured, objective measures of clinical reasoning and
decision-making skills. As a validated tool for assessment of
problem-solving skills, the SCT described in this study dem-
strates the potential to measure students’ decision-making
ability early in their medical education. The SCT shows
promise as a measure of “problem-solving” performance in
competency evaluation utilizing an efficient format that
approximates some of the ambiguities encountered in medical
practice. Our future plans include (1) an expansion of the SCT
administration to include more medical schools; (2) the
development of a database of validated questions from
which a subset can be drawn for optimization of longitudinal
test administration; and (3) a formal comparison of comput-
erized versus paper-based test administration for the SCT.
Acknowledgments

The authors thank Bernard Charlin who provided valuable comments and insight for the development and validation of the pre-clinical SCT. The authors also recognize the members of the Problem Solving Task Force at SBUMC and the Statewide Competency Assessment Taskforce at IUSM who made significant contributions to item development for the SCT. The support of the curriculum offices of IUSM and SBUMC was essential for the administration of the SCT at each site and we thank the key staff members involved in that effort. This research was conducted at the Indiana University School of Medicine and the Stony Brook University Medical Center.

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

Notes on contributors

ALOYSIUS J. HUMBERT, MD, is an assistant professor, Department of Emergency Medicine, Problem Solving Competency director, Indiana University School of Medicine.

MARY T. JOHNSON, PhD, is a professor, Department of Biomedical Sciences, assistant dean for Faculty Development, Florida State University College of Medicine.

EDWARD MIECH, EdD, is a research scientist, Health Services Research and Development, Roudebush Veteran Affairs Medical Center.

FRED FRIEDBERG, PhD, is an associate professor, Department of Psychiatry, Stony Brook University Medical Center.

JANICE A. GRACKIN, PhD, is an associate dean for Academic Affairs and Assessment, Nassau Community College.

PEGGY A. SEIDMAN, MD, is an associate professor, Department of Anesthesiology and Pediatrics, chair, Problem Solving Task Force, Stony Brook University Medical Center.

References


