RESEARCH BASIC TO MEDICAL EDUCATION

The Script Concordance Test: A Tool to Assess the Reflective Clinician

Bernard Charlin
Faculty of Medicine
University of Montreal
Quebec, Canada

Carlos Brailovsky
Faculty of Medicine
Laval University
Quebec City, Quebec, Canada

Louise Roy
Department of Family Medicine
University of Montreal
Quebec, Canada

François Goulet
Practice Enhancement Division
Collège des Médecins du Québec
Montreal, Canada

Cees van der Vleuten
Department of Educational Development and Research
University of Maastricht
Maastricht, The Netherlands

Background: The Script Concordance (SC) test is a new assessment tool. It is designed to probe whether knowledge of examinees is efficiently organized for clinical actions. That kind of organization of knowledge is named a script. The SC test places examinees in written, but authentic, clinical situations in which they must interpret data to make decisions. Purpose: The SC test is designed to measure the degree of concordance that exists between examinees 'scripts and scripts of a panel of experts. The objective of this article is to provide interested educators with the practical "how to " information needed to build and use an SC test.

Methods: The theoretical background of the SC test is described. The principles of construction of an SC test are presented, including the writing of clinical cases, the choice of item format, the validation of the test, and the elaboration of the scoring system. Results: A series of studies have shown that the SC test has interesting psychometric properties, in terms of reliability, face validity, and construct validity. Results from these studies are succinctly presented and commented. Conclusion: The SC test is a simple and direct approach to testing organization and use of knowledge. It has the strong advantage for a testing method of being relatively easy to construct and use and to be machine-scorable. It can be either paper- or computer-based and can be used in undergraduate, postgraduate, or continuing medical education.

Professional practice addresses problems that do not always have straightforward, algorithmic solutions. At the core of professional competence are judgment and insight, which rest on tacit knowledge. That kind of knowledge is neither visible nor tangible, and it cannot be evaluated easily using multiple-choice questions; yet, it is the touchstone of competent professional practice.' It is revealed only in action, in authentic situations when practitioners have to reflect on real concerns.[2-4]

Several authors[5-8] hypothesized that, in clinical medicine, skilled and experienced practitioners differ from those less experienced and skilled because they possess elaborated networks of knowledge fitted to their regular tasks. These networks, named scripts,[9-11] are organized to fulfill goals within tasks concerning

diagnosis, strategies of investigation, or treatment options. They begin to appear when students are faced with their first clinical cases and are then developed and refined during the whole clinical life.lo,t t

In this article, we describe a new assessment tool, the Script Concordance (SC) test, that stems from this cognitive theory of clinical expertise development. It places examinees in written but authentic clinical situations in which they have to interpret data to make decisions. The test belongs to the class of written simulations, [12] which could be either paper or computer based. It can be used in undergraduate, postgraduate, or continuing medical education. Our goal is to provide interested educators with all the information needed to construct and use the SC test, adapted to their needs.

Theoretical Background

In cognitive research on medical expertise, there has been a shift from the search for a generic problem-solving skill toward a focus on memory organization, knowledge use, problem representation, and how they change with experience. In the testing and evaluation domain, this change of focus has not had many applications. In this perspective, Elstein et a1.[13] suggested that evaluation should concentrate on judging the quality of a set of cognitive operations or knowledge structures by comparing a student's problem representation, judgments, and choices to those of the experienced group.

The SC test follows this approach. It is based on the script theory,[9-11] which postulates that in specific situations clinicians mobilize prestored sets of knowledge (their scripts) that are used to understand the situation and act according to specific goals (e.g., diagnosis, investigation, or treatment). Scripts of experienced clinicians vary on details, because each clinician has his or her own clinical experience, but they are similar for the essential elements. If it were not the case, clinicians would be unable to communicate efficiently about diseases or patients, and they would not reach the same diagnosis in similar situations. Scripts contain

information about the links that unite the items of knowledge (clinical features) related with an illness. It is these links, in diagnosis situations, that allow a person to make decisions concerning the strength or the weakness of a hypothesis or to decide if a clinical feature is never associated with such a hypothesis, in which case the hypothesis has to be rejected. Similar links are used to manage investigation or treatment decisions. tl

The test approach consists of presenting examinees with a series of patient problems and then asking examinees to make diagnostic, investigative, or therapeutic decisions when specific elements of information are provided. Examples of test items are given in Table 1. The test is designed to probe whether the organization of clinical knowledge (i.e., whether the nature of the links between items of knowledge) allows adequate clinical decisions. The test intends to assess the meaningfulness of the links among items, rather than assessing items in isolation. The scoring system of the test is designed to measure the distance, or the gap, that exists between examinees' scripts and scripts of a panel of experts.

Principles of Construction of the SC Test

Construction requires the collaboration of a small number of experts (two is usually sufficient at the stage of test item production). They are asked in an informal interview to describe some clinical situations that are representative of the field and are problematic. They then must specify for each situation (a) the relevant hypotheses, investigation strategies, or treatment options; (b) the questions they ask, physical examinations they perform, and tests they order to solve the problem; and (c) what clinical information, positive or negative, they would look for in these inquiries. In an SC test, there is no need to look for unusual clinical data. It is possible to discriminate among examinees with common data that require interpretation. Test items are built using the material obtained at this stage. The test consists of

Table 1. Example of Items From the Diagnostic Section of a Test

Clinical Vignette:	Joyce, 20 years old, is consulting at your office for a "vaginal discharge" she has been experiencing for the past week. She has had a new sexual partner for the past three months and she is worried about getting a sexually transmitted disease.

If You Were Thinking of (Infection)	And Then the Patient Reports or You Find on Clinical Examination	This Hypothesis **Becomes**				
Yeast	She had a sexually transmitted disease a few years ago	-2	-1	0	+1	+2
Chlamydia	She is taking a contraceptive pill	-2	-1	0	+1	+2
Herpes	She has an itchy vulvae	-2	-1	0	+1	+2
Herpes	She has dysuria	-2	-1	0	+1	+2
Yeast	Her discharge is greenish and itchy	-2	-1	0	+1	+2

Note: -2 =ruled out or almost ruled out; -1 =less probable; 0= neither less nor more probable; +1 =more probable; +2 =certain or almost certain.

several patient problems, presented in short vignettes, each of them followed by a series of related test items.

Clinical Vignettes

Each part of the test is based on a clinical case described in a few sentences. The description could be simple, as in "the patient is in her third semester of pregnancy and she is bleeding" for a diagnostic knowledge assessment, or it could be more detailed, as in the following for an assessment of treatment in coronary vascular disease:

> A 50-year-old woman presents a history of documented angina pectoris. She is sedentary but has no other risk factors. She had a hysterectomy with bilateral ovariectomy at the age of 39 and has never had replacement hormonotherapy, although there was no contra-indication for its use.

The description must contain systematically all of the necessary information for an expert to make an informed choice in the situation. For instance, in a situation of therapeutic choice in cardiac failure, the expert must know if the patient is short of breath and has abnormal pulmonary rates before deciding if a diuretic will be prescribed.

In some assessment situations, there is a need to describe an evolution of the situation. For instance, in a geriatric assessment tool, the first set of information could be similar to the following: "An 82-year-old woman is howling in the emergency room and pulls out the solution that contains the antibiotics required by her pneumonia. The attending staff has given a diagnosis of delirium whose origin has to be found." After the presentation of a series of diagnosis questions, a text describes the new requirement of the situation: "The nurse now asks for an efficient treatment for the delirium while waiting for the results of the investigation." This information is followed by a new series of treatment-related questions.

Choice of Test Item Format

The item format differs with the objective of assessment (diagnosis, investigation, or treatment), and for a given vignette, items are regrouped by format (e.g., some items on diagnosis, followed by some items on investigation). Each test item consists of three parts. The first part includes a diagnostic hypothesis, an investigative action, or a treatment option that is relevant to the situation. The second presents new information (e.g., a sign, condition, imaging study, or laboratory test result) that might have an effect on the diagnostic hypothesis, investigative action, or treatment option. The third part is a 5-point Likert-type scale. An illustration of the three formats is provided in Figure 1. Other formats may be established to assess other situations, such as giving a prognosis, or providing counseling.

Construction of Test Items

The construction of items follows the key features approach;[14] that is, the choice of question is focused on the elements that are the most useful to solve a clinical problem. Each item is built so that a reflection is necessary to answer it, and each is independent of the others. To prevent examinees from considering data on several following questions as cumulative information about the patient, hypotheses or options change for each question. It is also clearly specified in instructions for participants that within vignettes, each item is independent of the others. The goal of each item is not to determine the additive effect of a series of clinical information elements but to determine the effect of an isolated item of clinical information on a hypothesis, action, or treatment option. An example of a diagnostic section of a test is provided in Table 1.

Elstein et al.[15] showed that clinicians do not entertain large numbers of hypotheses at the same time. They found that when this number exceeds five, clinicians feel a cognitive need to reformulate their hypotheses in more inclusive and less numerous diagnoses. For this reason, we think that the number of tested hypotheses should not exceed five, although there should be at least two (if there is only one, this is not a diagnostic problem). The exact number depends on the relevance of the hypotheses to the situation.

To prevent a cueing effect on examinees, items are constructed to disperse answers among all values of the Likert-type scale. The number of items necessary for a test depends on its goal. For a Continuing Medical Education (CME) pretest, where the goal is to activate participants' prior knowledge and induce reflection on the appropriateness of that knowledge, the number of necessary items will be minimal (usually 20-30). For a test that will have certification or promotion consequences, where reliability is a major issue, the number of necessary items will be higher and will depend on the size of the probed domain.

Validation of the Test

The test is then submitted to a group of experts. The same group will serve for the elaboration of the scoring system. The definition of experts depends on the assessment situation. For example, to build a test that will serve to assess the knowledge of residents in cardiology, experts often will be chosen among certified cardiology specialists. Alternatively, if the test is built to assess knowledge for a CME activity aimed at family physicians, experts might be family physicians who

1-For ditutoostic knowledge assessment

If you were thinking of	And then you find	This hypothesis becomes
(a diagnostic hypothesis	*(a new clinical information, an imaging study or a laboratory test result)*	-2 -1 0 +1 +2

-2   Ruled out or almost ruled out
-1   Less probable
 0   neither less or more probable
+1   More probable
+2   Certain or almost certain

2- For investigation knowledge assessment

If you were considering to ask	And then you find	This investigation becomes
(a diagnostic test)	*(a new clinical information an imaging study or a laboratory test result)*	-2 -1 0 +1 +2

-2   Contra-mdhcated totally oralmost totally
-1   Not useful or even detrimental
 0   Nor less nor more useful
+1   Useful
+2   Absolutely necessary

3- For treatment knowledge assessment

If you were considering to Prescribe	And then you find	That prescription becomes
(a treatment option)	*(a new clinical information. an imaging study or a laboratory test result)*	-2 -1 0 +1 +2

-2   Contra-indicated totally or almost totally
-1   Not useful or even detrimental
 0   Nor less nor more useful
+1   Useful
+2   Necessary or absolutely necessary

Figure 1. The item format varies with the object of assessment (e.g., diagnosis, investigation, treatment).

have an important part of their clinical activities within that domain, along with some cardiology specialists. During their completion of the test, experts are asked to identify the items they find confusing or not relevant. These items are then either discarded or rewritten.

Elaboration of the Scoring System

The number of experts used to develop the scoring system must be sufficient (5-10) to express the variability in answers that experts may show for each item. Our first studieslb[-18] showed that experts provide the same answer on some items but also provide different answers for others. This is in accordance with other studies, which showed that experts' answers vary when they have to solve problems, even in their own field of expertise. [19,20] The scoring process of the test is based on the principle that any expert answer reflects the opinion of an expert, and those answers for which there is no agreement among all the experts should not be discarded. In other words, any answer given by an expert has an intrinsic value, even if other experts do

not agree with it. Hence, scores for each item are computed from the frequencies given to each point of the Likert-type scale by the experts. Table 2 provides an example of the scoring grid obtained for the items of Table 1 by a set of 10 experts.

For the first item, eight experts answered (0), one expert answered (-1), and one expert answered (+1). Hence scores for a student who answers (-1) is 0.1, (0) is 0.8, and (+1) is 0.1. Other answers are scored 0. Items in an SC test do not have the same maximum value: For the first item the maximum score is 0.8, and for the third it is 0.5. That value depends on the agreement between experts. Scoring is weighted by the degree of agreement between experts. This weighting is in no way artificial or arbitrary; it reflects the way experts answer the question.

The results of the test are represented by the sum of the scores obtained at each item. The maximum score for a test is the sum of the higher score obtainable on each item. The total score for the set of 20 items from which Tables 1 and 3 are taken is 11.4. For the convenience of interpretation, it is suggested to transform all scores to get a maximum score of 100. A score of 100

Table 2. *Example of the Scoring Grid Obtained for Items With a Set of 10 Experts*

Clinical Vignette: Joyce, 20 years old, is consulting at your office for a `vaginal discharge" she has been experiencing for the past week.

She has had a new sexual partner for the past 3 months, and she is worried about getting a sexually transmitted disease.

If You Were Thinking of (Infection)	And Then the Patient Reports or You Find on Clinical Examination	This Hypothesis Becomes				
		-.2	-1	0	+1	*2
Yeast	She had a sexually transmitted disease a few years ago	0	0.1	0.8	0.1	0
Chlamydia	She is taking a contraceptive pill	0	0.1	0.8	0	0.1
Herpes	She has an itchy vulvae	0	0.5	0.1	0.4	0
Herpes	She has dysuria	0.1	0.1	0.2	0.5	0.1
Yeast	Her discharge is greenish and itchy	0.1	0.5	0.1	0.3	0

Note: The group was composed of general practitioners.

sianifies that the examinee gives on each item the answer that most experts provide, and the lower the score thc farther examinees are from the experts' prototypic , ipt for the situation.

Results From Previous Studies

Inprevious studies, three SC tests with different contvts (gynecology, radiology, and surgery) were admini),~red to different groups of participants. [16-18] Results from these studies are succinctly presented in Table 3 with mean scores, standard deviations, and size of groups. The students have shown the widest variability intheir scores, followed by the residents in the studies in which they took part (they were not part of the study done in surgery). Levene's test of homogeneity of variance was used to verify whether the variances were equal in each study. The results indicated that the variances could be considered equal. The factorial analysis of variance used to test the mean group differences has shown significant differences between students, residents, and faculty groups, as indicated in Table 3.

The scores increased with the clinical expertise of group participants, with the students receiving lower scores than the residents and the residents receiving lower scores than the faculty participants. These observations were similar in the three studies already performed.

We used generalizability studies to evaluate the internal consistency of each test administration. It was done with Etudgen for Macintosh.[21] Coefficients were calculated with the results of the group of students that participated in each study. The observed generalizability coefficients (identical to the coefficient alpha) for each test are presented in Table 4. D studies showed the number of items that are necessary in each test administration to achieve an alpha of 0.8.

Discussion

The principle of the SC test is to compare the script of examinees to those of experienced clinicians using a series of clinical tasks, in specific contexts. The test possesses several advantages related to reli

Table 3. *Comparison of Main Scores by Groups in Three Different Studies*

		M	*SD*	Size
Gynecology*	Faculty	78.1	7.2	15
	Residents	75.9	12.8	12
	Clerks	67.1	14.0	76
Radiology**	Faculty	86.7	7.0	6
	Residents	69.3	7.6	10
	Clerks	62.6	14.0	14
Surgery***	Faculty	80.5	8.8	9
	Clerks	62.1	9.5	66

*Clerks versus faculty, $p<.001$. Other comparisons between groups were not significant. **The three comparisons were $p <.00$ 1. ***$p < .001$, Welch analysis of variance and Bonferroni post hoc correction.

Table 4. *Number of Items in Three Different Studies*

	No. of Items	A-Observed	No. of Items for A x 0.8'
Gynecology	50	0.794	51
Radiology	48	0.804	46
Surgery	26	0.544	59

'Number of items, observed alpha, and number of items that are necessary to obtain an alpha value of 0.8 as calculated with a generalizability D study.

ability and validity issues, the scoring process, and its educational effects.

The reliability studies have shown good alpha coefficients around 0.8 in two SC test administrations, thus indicating homogeneity of items. In the case of the surgery study, where the number of items was low, D studies indicate that 50-60 items are sufficient to achieve alphas of 0.8. The discrimin { ation power of individual items was good.

In most research concerning assessment of competence, experienced clinicians score little better or even worse than end-of-training residents, although one would expect that greater experience would be reflected in scores.[12] This counterintuitive finding, called "the intermediate effect," indicates that the proxy measures used in many studies, especially multiple-choice tests, are probably measuring competence

poorly and are invalid indicators of the work clinicians actually do in the practice setting. [1] In contrast to these findings, in our first studies [t--][18] we found that with SC tests, scores increase with clinical experience (see Table 3). This lends support to the construct validity of the tool. An explanation might be that most assessment tools, especially written ones, probe factual knowledge and interpretation of data using factual knowledge. SC tests go further; they explore the capacity of data interpretation in the making of clinical decisions, clearly a skill that belongs more to clinical competence than the simple recall of factual data.

Another measure of test validity is the face validity and the relevance of the tasks posed to examinees. In SC tests, examinees must solve problems belonging to the real world of the profession and must answer questions that experts consider of crucial importance in the process of solving that problem. Advantages of bringing these relevant contexts into test items are twofold. First, examinees (students, residents, or practicing physicians) find the test relevant and interesting to complete. Second, placing examinees in real-world situations allows expertise to emerge as the disappearance of the "intermediate effect" suggests.

The scoring system has several advantages. First, once the response grid is built, there is no answer interpretation, the test is standardized, scoring is straightforward, and the test is machine-scorable. Second, knowing that there is no single "best answer," the test can be used in test-retest situations. Finally, the test can be used in situations where there is no consensus among experts, in the literature or in practice. Among advantages of the test, it also is important to mention that relatively modest resources are required to develop it.

It is well known that assessment has a strong impact on learning. Students adapt what they learn to what they believe will be tested. SC tests reflect professional reality and are problem solving oriented; hence, they should influence the adaptation of students' learning activities in that direction. Furthermore, in CME activities, a SC test used at the beginning of activities to assess prior knowledge produced better retention of knowledge both at the end of the process and 3 months after the activity.[22] We interpret that positive effect as an activation of the knowledge that is relevant for the educational activity, which allows participants to detect where their prior knowledge might be insufficient or inaccurate.

Conclusion

The SC test is a simple and direct approach to testing organization and use of knowledge. It has the strong advantage for a testing method of being relatively easy to construct and use as well as of being machine-scorable. It can be either paper or computer

based and can be used in undergraduate, postgraduate, or continuing, medical education. It also has several psychometric advantages and good face validity for assessment of clinical competence. It is well accepted by students, residents, and physicians, [6,17,18,22] because examinees find that the tasks they have to fulfill are closely linked to professional reality and because it does not assess trivial knowledge, but rather probes real-world clinical knowledge. More research to support these claims is warranted and will be conducted.

References

I. McGaghie WC. Evaluating competence for professional practice. In L Curry, JF Wegin (Eds.), *Educating professionals. Responding to new expectationsfor competence and accountabili*ty (pp. 224-f I). San Francisco: Jossey-Bass, 1993.

2. Harris I. New expectations for professional competence. In L Curry, JF Wegin (Eds.), *Educating professionals. Responding to new expectations for competence and accountability (pp.* 17-52). San Francisco: Jossey-Bass, 1993.

3. Schbn DA. *Educating the reflective practitioner: Toward a nei,r design for teaching and learning in the professions.* San Francisco: Jossey-Bass, 1987.

4. Epstein RM. Mindful practice. *Journal ofthe American Medical Association* 1999;282:833-9.

5. Custers E, Regehr G, Norman GR. Mental representations of medical diagnostic knowledge: A review. *Academic Medicine* 1496;71(October Suppl.):S55-61.

6. Bordage G. Elaborated knowledge: A key to successful diagnostic thinking. *Academic Medicine* 1994;69:883-5.

7. Feltovich PJ. Expertise: Reorganizing and refining knowledge for use. *Professions Education Research Notes* 1983;4:5-7.

8. Zeitz CM. Some concrete advantages of abstraction: How experts' representations facilitate reasoning. In PJ Feltovich, KM Ford, RR Hoffman (Eds.), *Expertise in context: Human and machine (pp. 43-65).* Menlo Park, CA: AAAI Press, 1997.

9. Feltovich PJ, Barrows HS. Issues of generality in medical problem solving. In HG Schmidt, ML De Volder (Eds.), *Tutorials in problem-based learning: A new direction in teaching the health professions* (128-142). Assen, The Netherlands: Van Gorcum, 1984.

10. Schmidt HG, Norman GR, Boshuizen HPA. A cognitive perspective on medical expertise: Theory and implications. *Academic Medicine* 1990;65:611-21.

11, Charlin B, Tardif J, Boshuizen HPA. Scripts and medical diagnostic knowledge: Theory and applications for clinical reasoning instruction and research. *Academic Medicine* 2000;75:182-90.

12. Van der Vleuten CPM. The assessment of professional competence: Development, research and practical implications. *Advances in Health Sciences Education* 1996;1:41-67.

13. Elstein AS, Shulman LS, Sprafka SA. Medical problem solving, a ten-year retrospective. *Evaluation and the Health Profession* 1990;13:5-36.

14. Bordage G, Page G. An alternative approach to PMPs: The "key features" concept. In IR Hart, RM Harden (Eds.), *Further developments in assessing clinical competence (pp.* 59-75). Montreal, Canada: Heal, 1987.

15. Elstein AS, Shulman LS, Sprafka SA. *Medical problem solving: An analysis of clinical reasoning.* Cambridge, MA: Harvard University Press, 1978.

16. Charlin B, Brailovsky CA, Leduc C, Blouin D. The diagnostic script questionnaire: A new tool to assess a specific dimension of clinical competence. *Advances in Health Sciences Education* 1998;3:51-8.

17. Charlin B, Brailovsky CA, Brazeau-Lamontagne L, Samson L, Leduc C. Script questionnaires: Their use for assessment of diagnostic knowledge in radiology. *Medical Teacher* 1998;20: 567-71.
18. Brailovsky C, Charlin B, Beausoleil S, Coté S, van der Vleuten C. Measurement of clinical reflective capacity early in training as a predictor ofclinical reasoning performance at the end of residency: An exploratory study on the script concordance test. *Medical Education* 2000, accepted for publication.
19. Elstein AS, Holzman GB, Ravitch MM, et al. Comparison of physicians' decision regarding estrogen replacement therapy for menopausal women and decisions derived from a decision analytic model. *American Journal of Medicine* 1986;80:246-58.
20. Norman GR, Tugwell P, Feightner JW, Muzzin LJ, Jacoby LL. Knowledge and clinical problem solving. *Medical Education* 1985;19:344-536.
21. McNicoll A, Brailovsky CA, Bertrand R, Cardinet J. EtudGen, programme pour l'analyse de la généralisabilité pour Macintosh [EtudGen, program for the analysis of the generalizability for Macintosh]. CESSiJL 1992, 1996, 1999. In D Bain, G Pini (Eds.), *Pour évaluer vos évaluations: La généralisabilité, mode d'emploi (p.* 51). Geneva, Switzerland: Centre de recherches psychopédagogiques, 1996.
22. Brailovsky C, Charlin B, Émond C, Maltais P. *Script questionnaire as a method of assessing clinical reasoning after educational programs.* Workshop at the Alliance for Continuing Medical Education's 24th Annual Conference, Atlanta, GA, January 29, 1999.