Measurement of clinical reflective capacity early in training as a predictor of clinical reasoning performance at the end of residency: an experimental study on the script concordance test

C Brailovsky,[1] *B Charlin,*[2] *S Beausoleil,*[3] *S Cote*[3] *& C Van der Vleuten*[4]

Background The script concordance (SC) test was conceived to measure knowledge organization, the presence of links between items of knowledge which allow for interpretation of data in clinical decision making situations. Earlier studies have shown that the SC test has good psychometric qualities and overcomes some of the limitations of simulation clinical testing. This study explores the predictive validity of the test.

Objectives To verify whether scores obtained by students at the end of clerkship predict their clinical reasoning performance at the end of residency.

Design Comparison of scores obtained on a SC test taken at the end of clerkship with those obtained 2 years later at the end of residency on two clinical reasoning assessments of known validity, called the short-answer management problems (SAMPs) and the simulated office orals (SOOs), and an objective structured clinical examination (OSCE) aimed at assessing hands-on skills and clinical reasoning. Data were treated by Pearson correlation analysis.

Subjects and setting A cohort of 24 students from a medical school in Quebec was followed up to the end of their residency in family medicine, completed in several schools across Quebec.

Results The observed Pearson correlation coefficients of the SC test were statistically significant (0·451, $P = 0{\cdot}013$; 0·447; $P = 0{\cdot}015$) when compared with the SAMPs and the SOOs, respectively. They were not statistically significant (0·340, $P = 0{\cdot}052$) when compared with the OSCE.

Conclusion The authors assumed that the richness of knowledge organization, as indicated by SC test scores, would predict part of the performance on the measures of clinical reasoning (SAMP and SOO), but would predict less well performance on the OSCE which measures both clinical skills and clinical reasoning. Data found in the study are coherent with this hypothesis. This is evidence in favour of the construct validity of the SC test. It also indicates that, in the context of certification assessment, if a candidate has shown good organization of clinical knowledge at an early point in training, it can be expected that he/she will show good organization at subsequent measurements of this kind of knowledge. This appears to be true even if the later measures bear on a wider clinical domain.

Keywords Clinical clerkship, standards; *clinical competence; cohort studies; *educational measurement, *standards; knowledge, attitudes, practice; Quebec; reproducibility of results.

Medical Education 2001;35:430–436

Introduction

In recent decades, the assessment of clinical reasoning has evolved through several stages and has undergone theoretical development. In the 1960s, attempts were made to measure clinical reasoning using written simulations. The most prominent example was the patient management problems (PMP) approach. A realistic patient problem was presented and the examinee

[1] Evaluation Center for Health Sciences, Laval University Medical School, Quebec, Canada
[2] Unit of Research and Development for Medical Education, Faculty of Medicine, University of Montreal, Quebec, Canada
[3] Department of Surgery, University of Sherbrooke, Quebec, Canada
[4] Department of Educational Development and Research, University of Maastricht, The Netherlands

Correspondence: C A Brailovsky, Professor and Director, Evaluation Center for Health Sciences, Laval University, Québec G1K 7P4, Canada

Key learning points

Script concordance (SC) tests measure the presence of links between items of knowledge which characterizes knowledge adapted for efficiency in clinical tasks.

Good knowledge organization shown at the end of clerkship predicts clinical reasoning performance on assessments taken 2 years later.

Scores at the end of clerkship tend to predict well the scores on tests which mainly measure clinical reasoning, and less well the scores on tests which measure clinical skills more than clinical reasoning.

Script concordance (SC) tests measure the presence of links between items of knowledge which characterizes knowledge adapted for efficiency in clinical tasks.

Good knowledge organization shown at the end of clerkship predicts clinical reasoning performance on assessments taken 2 years later.

Scores at the end of clerkship tend to predict well the scores on tests which mainly measure clinical reasoning, and less well the scores on tests which measure clinical skills more than clinical reasoning.

was required to select the most pertinent items of history, physical examination, and investigations from a list. These simulations rapidly became part of many examination programmes with the aim of measuring a generic capacity of problem solving. Nevertheless, consistent empirical findings appeared which cast doubt on this endeavour. First, a score derived from one problem was not predictive for a score on another problem; the ability to solve problems was dependent on the 'content specificity' of the problem. Secondly, experienced clinicians scored hardly better and sometimes worse than less experienced clinicians or students.[1,2] This has been called 'the intermediate effect'.[3]

As a reaction to the 'content specificity' phenomenon, a new direction was suggested in the mid-1980s. It was argued that any clinical problem has one or more essential elements crucial to the management of the problem. For assessment purposes the suggestion was made that assessment should be limited to key elements in order to save time for testing additional problems and to improve reliability. This has been called the 'key features approach'.[4] Despite this improvement, results still showed superior performance for clinicians at the end of their training compared with clinicians with several years of experience,[5] when one would expect experience to provide a strong advantage in a measure of clinical competence. The intermediate effect persisted.

Several authors hypothesize that, in clinical medicine, skilled and experienced practitioners differ from less experienced and less skilled practitioners because they possess elaborate networks of knowledge which fit the tasks they perform regularly. These networks, called 'scripts',[6–8] are organized to fulfil goals within tasks concerning diagnosis, strategies of investigation, or treatment options. This kind of knowledge is revealed only in action, in authentic situations when practitioners have to reflect on real concerns. From this theoretical background, a new written simulation tool, the script concordance (SC) test, was designed to measure the richness of these networks. The principles of the SC test are described in another article.[9] The test places examinees in problematic clinical situations in which they must answer a question that experts ask in those situations and they have to interpret data to make decisions. Each test item has three parts (see Table 1). For diagnostic assessment items, the first part concerns a diagnostic hypothesis. The second is new informa-

Table 1 Sample of the script concordance test

You are thinking of the following hypothesis	*And you find*	*It has the following effect** (Please encircle your answer)						
1. Breast cancer	A patient more than 50 years old	A	B	C	D	E	F	G
2. Fibroadenoma	A patient younger than 30 years	A	B	C	D	E	F	G
3. Fibroadenoma	A very mobile breast mass	A	B	C	D	E	F	G
4. Cyst	An important inflammatory reaction	A	B	C	D	E	F	G
5. Cystic illness	Bilateral mass	A	B	C	D	E	F	G

*A = It can only be that hypothesis;
D = There is no effect on the hypothesis;
G = It definitely rejects the hypothesis.

tion, a sign or a symptom option which is relevant to the situation. The third part rates the decision made by the examinee on a Likert scale.

The choice of questions to be included in the test follow the key-feature approach, i.e. they are the questions experts think are the most likely to help solve the case. The scoring process involves weighting of responses through matching of expert judgements, and is an original part of the test. Several experts are asked to complete the test to indicate the most relevant answers. The scoring grid is based on their answers. Scores are computed from the frequencies given to each point of the Likert scale by the experts without any requirement to reach a consensus on a single answer. Hence the test measures the concordance between examinees' scripts and scripts of a panel of experts. It probes the organization of clinical knowledge, being intended to verify whether the knowledge of examinees, as described in Bordage's[10] classification of knowledge organization, is elaborated rather than dispersed. In order to do this, the test assesses how items of knowledge are structured and connected, including the nature of associations among items of knowledge, rather than the accumulation of items. With these innovations, we found that when clinicians were measured using this assessment tool the scores reflected their level of competence and experience.[11–13] The intermediate effect was no longer found and expertise level covaried systematically with an increase in performance.

In the present study we wanted to go further and to verify the predictive validity of the SC test. We asked all students who were completing their MD programme in a school of medicine to take an SC test. We then followed the cohort amongst them which entered a residency in family medicine. As candidates for licensure in family medicine, 2 years later, they were required to take a comprehensive licensure examination to confirm that they had acquired all the skills necessary to practise as independent clinicians. This licensure examination covers the complete domain of family/general practice and is comprised of three different tests.[14] The first one is a written test, called the short-answer management problems (SAMP) test, which follows the key-feature approach. Its validity and reliability are well known.[15] The SAMP test is composed of 42 clinical vignettes, each of which is followed by a series of three to five open-ended questions for which candidates provide written responses. The test measures the clinical reasoning skills required to make investigation, diagnosis, treatment or follow-up decisions.[15] The clinical vignettes require clinical reasoning in order to make decisions.[16] Because an adequate organization of knowledge is a prerequisite of problem-solving capacity, we expected that the SC test would correlate well with and be predictive of scores in the SAMP, a measure of problem-solving capacity.

The second test is an objective structured clinical examination (OSCE) at the end of the residency training. The goal of this OSCE is to assess clinical skills.[17–19] The third part of the licensure examination is another live standardized simulation. It is composed of five simulated office orals (SOOs).[20] Each SOO consists of a 15-minute interview during which the patient's role is played by a family physician who also scores the candidate's performance using a pre-tested objective marking scheme. Cases do not require a physical examination. The SOO assesses the candidates' abilities to manage complex biopsychosocial problems, with an emphasis on the patient–doctor relationship. The cases are based on the patient-centred method and the quality of clinical reasoning is essential for solving the problems presented.[21]

By virtue of test construction the SC test, the SOO and the SAMP (at least in part) measure output (SAMP) and output and process (SOO) of clinical reasoning, while the OSCE measures hands-on clinical skills as well as clinical reasoning. In the context of the multi-trait multi-method (MTMM) approach[22] we would predict that the homo-trait hetero-method approach (different methods measuring the same trait) should show good concordance between the SC test and the SOO test. Therefore we predicted in the study that correlations between the SC test, the SAMP and the SOO would be higher than the correlation between the SC test and the OSCE, despite the fact that the SAMP, the OSCE, and the SOO were taken at the same time, while the SC test had been taken 2 years earlier.

Materials and methods

Setting and subjects

In the school in which the study took place, the class who completed the clerkship in 1996 was composed of 90 students. At the end of the clerkship the whole class had a refresher course on essential concepts in surgery. Just before that, students were asked to take a script concordance test in surgery. The clinical areas which were used for the study were breast lump, gastrointestinal bleeding, acute abdominal pain and lump in the thyroid gland. The original SC test was composed of 60 items. After a first validation of the tool, 38 items distributed in the four content areas listed above were chosen, based on experts' and students' comments, and validated again with a group of experts who confirmed the quality of the selected items for assessment purposes. For the reliability analyses, the item scores were grouped by content areas and the grouped scores were used as input for reliability estimates.

A total of 66 students volunteered to take the SC test, among whom 24 had been accepted to a family medicine residency programme. These 24 students, 2 years later, took the licensure examination, which is mandatory in the Province of Quebec. The OSCE was composed of 26 clinical cases assessing different skills related to the practice of family medicine. The total weight of history taking, physical examination, organizational and communication skills represents about 40%, whereas investigation, diagnostic, treatment and follow up represent 60% of the exam, respectively. The SAMP test was composed of 42 clinical vignettes covering the whole field of family medicine. The simulated office orals included five cases. The scoring of the SOO is based on the fact that in the course of a good interview, a physician constantly weaves back and forth gathering information on both disease and illness. During this process, the definitions of disease and illness are integrated in order to arrive at a general understanding. For each SOO, candidates are scored on six items grouped in four areas: problem identification, social and development context, problem management and interview process and organization. The total score ranges from 6 to 36 and then is transformed into a percentage score.

Analysis

To evaluate whether the 66 clerkship students who volunteered to participate in the study were different from the 34 who declined participation, we compared the scores of the two groups on a comprehensive written examination. This examination was mandatory for clerks in order to obtain their MD degree. There were no significant differences between the two groups.

In order to judge the validity coefficients, reliabilities were estimated using generalizability analyses. The other statistics were descriptive statistics which consisted of the candidates' mean scores and standard deviations on each examination tool, and correlation studies assessing the association between scores on each measurement tool, with the corresponding tests of statistical significance. In the case of the SC test coefficient we added the item scores within content area scores ($n = 4$), and we used these four new scores as input for the reliability studies in order to avoid boosting of intercorrelations due to possible content connections.

It is well known that the reliability of a test affects its validity because, theoretically, a test cannot correlate more highly with any other score than it correlates with its own true score.[21] An estimate of the correlation that would be obtained if the tests did not contain measurement error can be obtained with the correction for attenuation. The correction for attenuation is pertinent if a test is used to predict the score on another test. We have used the correction for attenuation to assess the potential predictive validity of the SC test to predict the scores of the three tools used in the licensure examination.

Results

Descriptive statistics

The means and standard deviations (SD) were as follows: SC test 62·1 (9·5); OSCE 69·8 (5·0); SAMP 69·1 (4·5), and SOO 67·3 (7·7). To get the highest possible score on the SC test, an examinee would have to provide the answer most frequently given by experts on each item.

It appeared to be important to compare the group candidates' performances to the performances of the whole class in order to estimate whether or not they showed similar characteristics. We used unpaired *t* test analyses and the results are shown in Table 2.

Reliability analyses

The generalizability studies were done with EtudGen for Macintosh.[24] The generalizability coefficients for each test are presented in bold in Table 3. The SC test coefficient was 0·544 ($n = 66$). This coefficient was obtained by adding the item scores within content area scores ($n = 4$), and we used these four new scores as input for the reliability studies in order to avoid boosting of the intercorrelations due to possible content connections. For the OSCE, the SAMP, and the SOO the generalizability (G) coefficients were 0·717 ($n = 181$), 0·817 ($n = 769$) and 0·478 ($n = 769$), respectively. The SOO showed strong item–total correlations (>0·478),

Table 2 Comparison of the scores of the sample group with the scores for the whole group for each part of the licensing exam

	Whole group	Sample (n = 24)	*t* value*	*P* value
OSCE	68·3	70·0	1·898	0·0591
SAMP	68·8	70·2	1·127	0·2610
SOO	66·5	69·4	1·821	0·0701

* Unpaired *t* test, hypothesized difference = 0.
OSCE, objective structured clinical examination; SAMP, short-answer management problems; SOO, simulated office orals.

Table 3 Correlations among the scores obtained by the students ($n = 24$) with the four assessment tools described in the study

	SC	OSCE	SAMP	SOO
Observed correlations				
SC		0·340	0·451	0·447
OSCE			0·449	0·426
SAMP				0·171
*True/corrected correlations**				
SC		0·544	0·667	0·729
OSCE			0·586	0·606
SAMP				0·228
Reliability coefficient	**0·544**	**0·717**	**0·817**	**0·478**

* Correlation coefficients corrected for attenuation.
SC, script concordance; OSCE, objective structured clinical examination; SAMP, short-answer management problems; SOO, simulated office orals.

thus supporting the construct that is being measured. The decision studies (*D*-studies) showed that nine SOO cases would have a *G* coefficient of 0·784.

Correlation studies

The observed Pearson correlation coefficient regarding the SC test as a predictor of SAMP scores was 0·451. It was statistically significant ($P = 0·013$). When corrected for attenuation due to the unreliability of the tests, the coefficient reaches a value of 0·667.

In the case of SOO scores, the correlation coefficient was 0·447. It was also statistically significant ($P = 0·015$), and when corrected for attenuation, the coefficient attained a value of 0·729.

The observed Pearson correlation coefficient concerning the SC test as a predictor of the OSCE was low ($r = 0·340$), and not statistically significant ($P = 0·052$). When corrected for attenuation, the coefficient attained a value of 0·544 (Table 3).

Discussion

These results show that the SC test, taken at the end of clerkship in the specific domain of surgery, effectively predicts the scores obtained 2 years later, at the end of residency, on tests of reasoning, even if they are related to a comprehensive assessment of clinical family medicine knowledge. They also show that the SC test predicts less well the scores on an OSCE which measures both hands-on skills and clinical reasoning. In this case, it is possible to argue that the observed scores on the OSCE are the sum of two different elements interacting within the performance measured.

The study has strengths and weaknesses. The assessment instruments used at the end of residency have well-established validity as instruments of measurement of clinical reasoning (SAMP and SOO),[15,20] and of hands-on clinical skills and reasoning (OSCE),[17,19] with good reliability for the three. The reliability analyses were performed with the cohort of candidates which sat for each exam. As mentioned by Norman *et al.*[25] coefficients are very unstable with small sample sizes, and because reliabilities appear in the denominator of the formula for calculating the true correlation, it is important to have good reliability estimations. The SC test is an instrument which is under development, but preliminary studies have shown good reliability and there are arguments in favour of its construct validity.[11–13] The results of this study concern a cohort of only 24 students who completed a residency in family medicine. This is not a large sample, but it was sufficient to detect an effect in the study. The two series of measures are 2 years apart in time. This is a long period in a learning environment, and many confounding variables can have an impact on students' learning. However there is always much 'noise' in educational measurement, and we can postulate that the impact of these confounding variables may be found to be equally distributed among the observed scores of the three tests and could explain the results. Despite the noise and the 2-year time interval we still observe a relatively strong correlation among the variables under study.

The SC test is designed in the context of a theoretical background, the script theory, which states that the development of clinical reasoning competence requires a reorganization of knowledge in order to be able to fulfil the specific demands of clinical tasks, mainly diagnosis, investigation and treatment. Bordage[10] has shown that some clinicians organize their knowledge well and that others do not. The goal of the SC test is to identify the students who have organized their knowledge for efficient use in their clinical work, those whose knowledge reaches an elaborated structure. The results of this study furnish evidence in favour of the construct validity of the SC test. One explanation for these results is that good clinicians structure their knowledge in order to fulfil their clinical tasks, while others only accumulate knowledge without transforming it to fit the tasks. This is in accordance with the developmental theory of clinical competence of Schmidt, Norman and Boshuizen.[7] When students are confronted with their first clinical cases, they must restructure their knowledge to their new tasks. Possibly some students do not perform this restructuring, or carry it out incompletely.

This study suggests that, in the context of certification assessment, if a candidate has shown good organization of clinical knowledge at a particular time during training, it can be expected that the candidate will show good organization at subsequent measurements of this kind of knowledge. This appears to be true even if the measures bear on a wider clinical domain.

If this is confirmed, the SC test may have applications in several fields of medical education, such as selection of candidates for residency or identification of students who have problems with knowledge organization and who need remedial training. It also has implications for other fields of study, because it indicates that if content specificity is a strong factor in clinical reasoning expertise, there is also another factor which predicts performance, that is, the organization of knowledge.

Contributors

C. Brailovsky, B. Charlin and C. Van der Vleuten developed the project and the research protocol, performed the analyses and wrote the paper. S. Beausoleil and S. Coté developed the script questionnaire under the supervision of B. Charlin.

Funding

This work received funding from the Medical Council of Canada and from the Medical Research Council of Canada/Association of Canadian Medical Colleges.

References

1 van der Vleuten CPM, Newble D, Case S, Holsgrove G, McCann B, McRae, *et al.* Methods of assessment in certification, In: Newble D, Jolly B, Wakeford R. eds. *The Certification and Recertification of Doctors: Issues in the Assessment of Clinical Competence.* Cambridge: Cambridge University Press. 1994.

2 van der Vleuten CPM, Luijk SJ, van Beckers HJM. A written test as an alternative to performance testing. *Med Educ* 1989;2**3**:97–107.

3 Schmidt HG, Boshuizen HPA, Hobus PPM. Transitory stages in the development of medical expertise: the 'intermediate effect' in clinical case representation studies. In: *Proceedings of the 10th Annual Conference of the Cognitive Science Society.* Hillsdale, NJ: Erlbaum; 1988.

4 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.* Montreal: Heal-Publications; 1987: pp. 59–75.

5 Bordage G, Brailovsky CA, Cohen T, Page G. Maintaining and enhancing key decision-making skills from graduation into practice: an exploratory study. In: AJJA Scherpbier, CPM van der Vleuten, J-J Rethans, eds. *Advances in Medical EducationI.* Dordrecht, The Netherlands: Kluwer Academic; 1996: pp. 128–30.

6 Feltovich PJ. Expertise: reorganizing and refining knowledge for use. *Professions Educ Res Notes* 1983;**4**:5–7.

7 Schmidt HG, Norman GR, Boshuizen HPA. A cognitive perspective on medical expertise: theory and implications. *Acad Med* 1990;**65**:611–21.

8 Charlin B, Tardif J, Boshuizen HPA. Scripts and medical diagnostic knowledge: theory and applications for clinical reasoning instruction and research. *Acad Med* 2000;7**5**: 182–90.

9 Charlin B, Roy L, Brailovsky CA, Goulet F, Van der Vleuten CPM. The Script Concordance Test: A Tool to Assess the Reflective Clinician. *Teaching Learning Med* 2000;**12**:183–8.

10 Bordage G. Elaborated knowledge: a key to successful diagnostic thinking. *Acad Med* 1994;6**9**:883–5.

11 Charlin B, Brailovsky CA, Brazeau-Lamontagne L, Samson L, Leduc C. Script questionnaires: their use for assessment of diagnostic knowledge in radiology. *Med Teacher* 1998;2**0**:567–71.

12 Charlin B, Brailovsky CA, Leduc C, Blouin D. The diagnostic script questionnaire: a new tool to assess a specific dimension of clinical competence. *Adv Health Sci Educ* 1998;**3**:51–8.

13 Brailovsky C, Charlin B, Émond JG, Miller F, Maltais P. Script questionnaire as a method of assessing clinical reasoning after educational programs. Workshop. *Alliance for Continuing Medical Education's 24th Annual Conference;* 1999 Jan 29; Atlanta, Georgia.

14 Grand'Maison P, Lescop J, Rainsberry P, Brailovsky CA. Large-scale use of an objective, structured clinical examination for licensing family physicians. *Can Med Assoc J* 1992;**146**:1735–40.

15 Handfield-Jones R, Brown JB, Biehn JB, Rainsberry P, Brailovsky CA. Certification Examination of the College of Family Physicians of Canada. Part 3: short answer management problems. *Can Fam Physician* 1996;4**2**:1353–61.

16 Anonymous. Internal report. Toronto: College of Family Physicians of Canada; 2000.

17 Grand'Maison P, Lescop J, Brailovsky CA. Large-scale use of an objective, structured clinical examination for licensing family physicians. *Can Med Assoc J* 1992;4**6**:1735–40.

18 Brailovsky CA, Grand'Maison P, Lescop JA. A large-scale multicenter Objective Structured Clinical Examination for licensure. *Acad Med* 1992;67:S37–9.

19 Brailovsky CA, Grand'Maison P. A Using evidence to improve evaluation: a comprehensive psychometric assessment of a SP-based OSCE licensing examination. *Adv Heath Sci Educ* 2000;**5**:207–19.

20 Handfield-Jones R, Brown JB, Rainsberry P, Brailovsky CA. The Certification Examination of the College of Family Physicians of Canada: (IV) simulated office orals. *Can Fam Physician* 1996;4**2**:1539–48.

21 Handfield-Jones R, Brown JB, Rainsberry P, Brailovsky CA. The Certification Examination of the College of Family

Physicians of Canada: (II) conduct and general performance. *Can Fam Physician* 1996;**42**:1188–95.

22 Campbell DT, Fiske DW. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychol Bull* 1959;**54**:181–05.

23 Allen MJ, Yen WM. *Introduction to Measurement Theory.* Belmont, California: Wadsworth; 1979.

24 McNicoll A, Brailovsky CA, Bertrand R, Cardinet J. Etud-Gen, programme pour l'analyse de la généralisabilité pour Macintosh. In: D Bain, G Pini, eds. *Pour Évaluer Vos Évaluations: La Généralisabilité, Mode D'emploi.* Geneva: Centre de Recherches Psychopédagogiques; 1996.

25 Norman GR, Swanson DB, Case SM. Conceptual and methodological issues in studies comparing assessment formats. *Teaching Learning Med* 1996;**8**:208–16.

Received 7 *March* 2000; *editorial comments to authors* 7 *June* 2000; *accepted for publication* 22 *July* 2000