Medical Teacher, Vol. 24, No. 5, 2002, pp. 522–527

Taylor&Francis healthsciences

Stability of clinical reasoning assessment results with the Script Concordance test across two different linguistic, cultural and learning environments

LOUIS SIBERT[1], BERNARD CHARLIN[2], JACQUES CORCOS[3], ROBERT GAGNON[2], PHILIPPE GRISE[1] & CEES VAN DER VLEUTEN[4]

[1]Department of Urology, Rouen University Hospital, Rouen, France; [2]Unit of Research and Development in Medical Education, Faculty of Medicine, University of Montreal, Montreal, Canada; [3]Department of Urology, Sir Mortimer B. Davis-Jewish General Hospital, McGill University, Montreal, Canada; [4]Department of Educational Development and Research, University of Maastricht, The Netherlands

SUMMARY *The Script Concordance (SC) test is designed to measure the organization of knowledge that allows interpretation of data in clinical reasoning. An originality of the test is that answer keys use an aggregate scoring method based on answers given by a panel of experts. Previous studies have shown that the SC test has good construct validity. This study, done in urology, explores (1) the stability of the construct validity of the test across two different linguistic and learning environments and (2) the effect of the use of experts who belong to different environments. An 80-item SC test was administered to participants from a French and a Canadian university. Two levels of experience were tested: 25 residents in urology (11 from the French university and 14 from the Canadian university) and 23 students (15 from the French faculty, eight from the Canadian faculty). Reliability analysis was studied with Cronbach's alpha coefficient. Scores between groups were compared by analysis of variance. Reliability coefficient of the 80 items test was 0.794 for the French participants and 0.795 for the Canadian participants. Scores increased with clinical experience in urology in the two sites. Candidates obtained higher scores when correction was done using the answer key provided by the experts from the same country. These data support the stability of the construct validity of the tool across different learning environments.*

Introduction

Until recently, the search for instruments to assess clinical competence was based on a traditional conception of the nature of clinical competence. The development of competence was considered as being equal to the development of each of its components. However, growth in competence appeared more capricious than expected. This empirical disillusion has stimulated cognitive psychological research into the development of expertise (Schmidt *et al.*, 1990; Van Der Vleuten, 1996). Several authors have established that, in most clinical situations, medical reasoning is a hypothetico-deductive process (Elstein *et al.*, 1978; Barrows *et al.*, 1982) characterized by early generation of hypotheses, oriented data collection and decision-making judgement, using collected data to confirm or reject hypotheses. Expertise development of professionals appears to be strongly related to knowledge. However, the way in which knowledge is stored, used and retrieved characterizes differences between novices and experts. Experienced practitioners possess elaborated networks of knowledge fitted to the tasks they regularly do (Feltovich, 1983; Custers *et al.*, 1996; Zeitz, 1997). These networks, named scripts (Feltovich, 1983; Schmidt *et al.*, 1990; Charlin *et al.*, 2000a), are organized to fulfil 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 later developed and refined during their entire clinical career (Schmidt *et al.*, 1990; Bordage, 1994). The script theory states that in diagnostic situations clinicians fill their working memory with knowledge related to each relevant hypothesis. This activated knowledge is then used in a deductive process to actively seek information that will allow confirmation or rejection of respective hypotheses (Charlin *et al.*, 2000a).

These cognitive psychological models have proved to be useful for new directions in assessment. A new written assessment tool, the Script Concordance (SC) test, was in fact designed to measure the richness of these networks (Charlin *et al.*, 2000b). 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. It places examinees in written but authentic clinical situations where they have to interpret data to make decisions.

Contemporary assessment tools of clinical reasoning have repeatedly shown the puzzling fact that experienced clinicians score hardly better and sometimes worse than less experienced clinicians or students (Van der Vleuten, 1996). This counterintuitive finding, called 'the intermediate effect', indicates that most of these methods, especially written examinations, i.e. multiple-choice tests, measure clinical factual knowledge rather than clinical reasoning competence and are invalid indicators of the work clinicians actually do in a practice setting (McGaghie, 1993). In contrast with these

Correspondence: Dr Louis Sibert, Department of Urology, Rouen University Hospital-Charles-Nicolle, Pavillon Derocque, 1 rue de Germont, 76031 Rouen Cedex, France. Tel: (33) 2 32 88 81 73; fax: (33) 2 32 88 82 05; email: louis.sibert@chu-rouen.fr

ISSN 0142–159X (print)/ISSN 1466–187X online/02/050522–06 © 2002 Taylor & Francis Ltd
DOI: 10.1080/0142159021000012599

findings, earlier studies have shown that SC tests have good construct validity with an increase in the mean scores of participants with different levels of clinical expertise (Charlin *et al.*, 1998, 2000b). The SC test was developed to explore the capacity of data interpretation when making clinical decisions, clearly a skill that belongs more to clinical competence than the simple recall of factual data.

The scoring process uses an aggregate scoring method (Norman, 1985; Norcini *et al.*, 1990). The principle is that any experienced clinician response is a reflection of expertise, and responses for which there is no agreement among experts should not be discarded. The test is submitted to a group of experienced clinicians and scoring is weighted by the degree of agreement between experts. This scoring is in no way artificial or arbitrary but it could be influenced by the cultural environment and the field of clinical activities of the reference panel.

In previous studies on the SC test (Charlin *et al.*, 1998, 2000b), candidates sat their examination in the same language (French) and were trained in one common medical culture. In the present study, we looked for arguments supporting the idea that the same test would be able to discriminate among participants across different countries according to their level of clinical experience, despite cultural, educational and institutional differences. The domain of assessment was urology. Two research hypotheses were tested: (1) the capacity of the test to discriminate among candidates with different levels of experience should persist even if the learning and cultural environment of the reference panel is different; (2) candidates should obtain higher scores when they are assessed by experts from their own culture.

Methods

Construction of the SC test

Two clinicians belonging to the French faculty and the Canadian faculty were asked to describe clinical situations representative of urology practice and based on major educational objectives of Canadian and French urology training programmes. They were asked to 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 request to solve the problem; and (c) what clinical information, positive or negative, they would look for in these inquiries. Test items were built with the material obtained from that inquiry. Great care was taken over the content validity of the test. The evaluation concerned urology residents as well as medical students, therefore all the areas of clinical competence (i.e. diagnosis, investigation, treatment) were assessed. Table 1 shows the blueprint of the test.

The clinical situations were presented in short vignettes, each of them followed by a series of related test items. Items were constructed according to the methodology described by Charlin *et al.* (2000b). The item format differs with the objective of assessment (diagnosis, investigation or treatment). Each item consists of three parts. The first part includes a diagnostic hypothesis, an investigative action or a treatment option. The second presents new information (e.g. clinical data, an imaging study or a laboratory test result) that might have an effect on the diagnostic hypothesis, investigative action or treatment option. The third part is a five-point Likert-type scale (see illustration of the three formats in Table 2). Each item was constructed so that reflection was necessary to answer it. Clear instructions were also given that all items within each vignette were independent from each other. Hypotheses or options change for each question. Hence, a test of 12 clinical situations and 85 items was then constituted (see example of items from the diagnostic section in Table 3).

The test was administered in English to the Canadian subjects and in French to the French participants. Great care was taken in the translation process in order to avoid word substitution, word omission, word addition or different meaning for a word in the other language (Ferland *et al.*, 1983). Each version of the test was then reviewed by experienced urologists in their own language. During their completion of the test, urologists were asked to identify the items they found confusing or not relevant. Five items were then discarded. Eighty items were retained for the calculation of scores and statistical analysis.

Table 1. Blueprint of the Script Concordance test in urology.

Clinical problems	Context	Age	Sex	NB	Assessed components
Scrotum enlargement	C	42	M	5	DG
Scrotal trauma	E	27	M	5	TT
Renal colic	E	63	F	10	DG, I
Pelvis trauma	E	25	M	10	DG, TT
Infertility	C	31	M	5	DG
Urinary incontinence	C	57	F	5	DG
Urinary lithiasis	E	35	M	10	DG, TT
BPH, Prostate cancer	C	58	M	10	I, TT
Erectile dysfunction	C	66	M	5	DG
Obstructive renal failure	E	73	F	5	TT
Urinary retention	E	71	M	5	TT
Kidney tumour	C	73	F	5	DG

Notes: NB = number of questions, C = Consultation, E = emergency, M = male, F = female, DG = diagnostic, I = investigation, TT = treatment, BPH = benign prostatic hypertrophy.

Table 2. Illustration of questions and answering grids format.

For diagnostic knowledge assessment[a]		
If you were thinking of	And then you find	This hypothesis becomes
(A diagnostic hypothesis)	(New clinical information, an imaging study or a laboratory test result)	-2 -1 0 +1 +2
For investigation knowledge assessment[b]		
If you were considering requesting	And then you find	This investigation becomes
(A diagnostic test)	(New clinical information, an imaging study or a laboratory test result)	-2 -1 0 +1 +2
For treatment knowledge assessment[c]		
If you were considering prescribing	And then you find	The relevance of this treatment becomes
(A therapeutic option)	(New clinical information, an imaging study or a laboratory test result)	-2 -1 0 +1 +2

Notes: The item format varies with the object of assessment (e.g. diagnostic, investigation, treatment).
[a] -2 = the hypothesis is almost eliminated; -1 = the hypothesis becomes less probable; 0 = the information has no effect on the hypothesis; +1 = the hypothesis is becoming more probable; +2 = it can only be this hypothesis.
[b] -2 = contra-indicated totally or almost totally; -1 = not useful or even detrimental; 0 = not less not more useful; +1 = useful; +2 = absolutely necessary.
[c] -2 = contra-indicated totally or almost totally; -1 = not useful or even detrimental; 0= neither less nor more useful; +1 = useful; +2 = necessary or absolutely necessary.

Table 3. Example of items from the diagnostic section of the SC test in urology.
Clinical vignette: A 25-year-old male patient is admitted to the emergency room after a fall from a motorcycle with a direct impact to the pubis. Vital signs are normal. The X-ray reveals a fracture of the pelvis with a disjunction of the pubic symphysis.

If you were thinking of	And then you find	This hypothesis becomes
Urethral rupture	Urethral bleeding	-2 -1 0 +1 +2
Retroperitoneal bladder rupture	Bladder distension	-2 -1 0 +1 +2
Urethral rupture	Upward and bulging prostatic apex at the digital rectal examination	-2 -1 0 +1 +2
Intra-peritoneal bladder rupture	Spontaneous micturition after the accident	-2 -1 0 +1 +2
Urethral rupture	Perineal haematoma	-2 -1 0 +1 +2

Notes : -2 = the hypothesis is almost eliminated; -1 = the hypothesis becomes less probable; 0 = the information has no effect on the hypothesis; +1 = the hypothesis is becoming more probable; +2 = it can only be this hypothesis.

Participants

The test was submitted to two groups of participants from the urology departments of McGill University, Montreal, Canada (14 residents, eight medical students), and Rouen University Hospital, France (11 residents, 15 medical students). Two groups of certified urologists made the reference panels: 10 French urologists and 12 Canadian urologists. French urologists were representative of different areas of practice (faculty members, general and private practices). Canadian urologists were exclusively faculty members. Each urologist had a minimum of five years' clinical experience. All the Canadian participants were English speaking. All subjects asked to participate volunteered. Criteria for inclusion were: for urologists, to be certified according to the rules of their country; for residents, to belong to the urology programme at McGill or Rouen University Hospital; and, for students, to have had a rotation in a urology department in the last six months. All students had spent from four to six months on urology attachments. In this exploratory research, the two groups of urologists were successively considered as the reference panel for the construction of answer keys.

Scoring process

For each item, answers were assigned a weight corresponding to the proportion of the members of the reference panel who selected it. Credits for each answer were then transformed proportionally (division of all scores by the modal value on the item) to obtain a maximum score of 1 for modal

experts' choice(s) on each item, other experts' choices receiving a partial credit. Answers not chosen by any experts received zero. For example, if on an item, six experts (out of 10) chose response +1, this choice received 1 point (6/10 x 10/6), if four experts chose response +2, this choice received 0.64 (4/10 x 10/6). The total score for the test is represented by the sum of the scores obtained for each item. French candidates were scored by the French reference panel and Canadians candidates were scored by Canadian reference panel. This procedure was later reversed.

Statistical analysis

Descriptive statistics of the participants' scores on the SC test were performed, followed by an univariate analysis of variance to test the differences between groups' means according to the composition of the reference panel. Variations of scores obtained by students and residents with the two different reference panels were compared with Student's *t*-test. The homogeneity of group variances was estimated with Levene's test in order to interpret the results of the previous analysis. When variances were unequal, an adapted test was used. To evaluate the presence of a significant statistical difference, a $p < 0.05$ value was considered as significant. Reliability of the examination was assessed through Cronbach's alpha internal consistency coefficient.

Results

The first analyses compared the scores obtained by candidates grouped by level of experience in urology, according to the composition of the reference panel. These results are summarized in Table 4. Mean global scores were 48.33 ± 5.64 for the students and 55.15 ± 4.21 for the residents when experts were the Canadian urologists. Mean global scores were 46.67 ± 5.60 for the students and 53.16 ± 4.68 for the residents when French urologists were considered as the reference panel. The univariate analysis of variance showed significant differences between students, residents and urologists. These observations were similar for the two different reference panels ($p < 0.0001$), indicating that the scores increased with the clinical experience of group participants, independently of the medical culture of the reference panel.

The second analyses compared the variations of candidates' scores according to their learning environment (see Table 5). Performance of French students did not vary for the two reference panels (46.51 ± 4.71 versus 46.95 ± 6.80, $p = 0.68$). Canadian students obtained higher scores from the Canadian reference panel (51.74 ± 5.95 versus 46.16 ± 2.33, $p < 0.0001$). French residents performed significantly better when French urologists were considered as the reference panel (56.18 ± 1.73 versus 54.62 ± 2.58, $p < 0.001$). Canadian residents' scores significantly increased when Canadian urologists were the experts (55.57 ± 5.22 versus 50.78 ± 4.94, $p < 0.0001$).

The Cronbach alpha reliability coefficients were similar for the French participants and the Canadian participants: 0.794 and 0.795 respectively.

Table 4. Comparison of mean scores by groups according to composition of reference panels.

Composition of the reference panel	Residents ($n = 25$)[a]	Students ($n = 23$)	p-value[b]
French urologists ($n = 10$)	53.16 ± 4.68[c]	46.67 ± 5.60	0.0001
Canadian urologists ($n = 12$)	55.15 ± 4.21	48.33 ± 5.64	0.0001

Notes: [a]Size of groups, [b]$p < 0.05$ was considered as significant (univariate analysis of variance); [c]values are mean ± SD.

Discussion

Cross-cultural research in the field of clinical competence assessment is not common (Marshall *et al.*, 1995). Rare existing data were obtained from bilingual medical communities but with candidates trained in one common medical culture (Brailovsky *et al.*, 2000). Our results showed variations of scores among groups of candidates with the same level of experience in urology, whatever the composition of the reference panel. Globally, Canadian candidates obtained better scores when they were assessed by the Canadian urologists, and French candidates performed better when French urologists were considered as the reference panel.

To translate the examination we followed the process as described by Marshall *et al.* (1995). This process was effective for minimizing translation errors and assuring equivalency of the English and French versions of the SC test. We believe that the differences observed are not biases of candidates' language but reflect medical and cultural differences between the two sites. Variance in urological practice patterns in the two communities influenced the performance of candidates in the SC test. In North America, urology practice is currently specialized and restricted to a specific field of urology. In comparison, French urologists have a general as well as specialized urology practices. For example, male infertility and urinary incontinence are very specialized patterns; these specific cases could be rare in some Canadian urologists' daily practice. In contrast, most of the French urologists are theoretically able to be effective in the initial care of these patients. Differences in areas of practice between the two panels of experts (Canadian urologists were exclusively faculty members, French urologists were from different areas

Table 5. Variations of mean scores according to learning environment of groups of candidates and reference panels.

Composition of reference panel	French students	Canadian students	French residents	Canadian residents
French urologists	46.95 ± 6.80[a]	46.16 ± 2.33	56.18 ± 1.73	50.78 ± 4.94
Canadian Urologists	46.51 ± 4.71	51.74 ± 5.95	54.62 ± 2.58	55.57 ± 5.22
p-value[b]	0.68	0.0001	0.001	0.0001

Notes: [a]Values are mean ± SD, [b]$p < 0.05$ was considered as significant (Student's *t*-test).

of practice) could also explain the different levels of familiarity with the problems depicted in the test. This could also influenced the performance of candidates.

However, despite the fact that students and residents are trained in two different medical cultures, their performances on the SC test were remarkably similar. Our results show an increase in the mean scores on the SC test of groups with different clinical experience, with the students receiving lower scores than the residents. These observations were similar in two sites with different languages and learning environments. This may mean that the SC test measures a dimension for which, as one should expect, experienced clinicians get better scores than less experienced subjects. This supports the construct validity of the instrument. The study also showed that this test developed in urology has a relatively high reliability. The consensual effect of this kind of examination, in terms of construct validity and reliability across learning environments, across cultures and ways of practice, is impressive and underlines the advantages of the SC test.

In the SC test, examinees have to answer questions concerning real medical problem solving that experts consider of crucial importance to the process. It is well known that assessment has a strong impact on learning. Students and residents adapt what they learn to what they believe will be tested (Friedman Ben-David, 2000). The SC test could reflect professional reality and is problem solving oriented; hence it should influence the adaptation of students' and residents' learning activities in that direction.

Our data demonstrate (1) that the SC test could assess candidates with good construct validity and reliability even if the reference panel belongs to a different medical culture; (2) scores are globally higher when the scoring system is established with a reference panel from the same environment as candidates.

The current need for medical training programme harmonization throughout European Union faculties favour the development of cross-cultural research in the field of clinical competence assessment. In this context, the SC test appears promising and warrants further investigation to confirm its value as a multilinguistic and multicultural assessment tool of clinical reasoning.

Conclusions

These results showed an increase in the mean scores on the SC test of groups with different clinical experience in urology. The same effect was observed in two different learning and cultural environments. The variations of scores between groups of candidates with the same clinical experience could reflect the cultural difference among the groups of experts. These data present another argument in favour of the construct validity of the tool and the stability of its psychometric properties when it is administered in different learning environments. Most of the national medical associations are currently sensitive to the need to harmonize training throughout the European Union. In this context, our findings warrant consideration and further research.

Practice points

- The study confirms the construct validity of the Script Concordance test in assessing clinical reasoning in urology when it is administered in different learning environments.

Acknowledgement

The authors thank Richard Medeiros for his valuable advice in editing the manuscript.

Notes on contributors

LOUIS SIBERT is a Urologist in the Department of Urology, Rouen University Hospital, Rouen, France. He is also a member of the Department of Medical Education, Faculty of Medicine, Rouen, France.

BERNARD CHARLIN is a Professor of Surgery. He is the Director of the Unit of Research and Development in Medical Education, Faculty of Medicine, University of Montreal, Montreal, Canada.

JACQUES CORCOS is a Urologist. He is Associate Professor, McGill University and Chief of the Department of Urology, Sir Mortimer B. Davis-Jewish General Hospital, Montreal, Canada.

ROBERT GAGNON is a methodologist and member of the unit of Research and Development in Medical Education, Faculty of Medicine, University of Montreal, Canada.

PHILIPPE GRISE is Professor of Urology and Chief of the Department of Urology, Rouen University Hospital, France.

CEES VAN DER VLEUTEN is Professor and Chair, Department of Educational Development and Research, University of Maastricht, The Netherlands.

References

BARROWS, H.S., NORMAN, G.R., NEUFELD, V.R. & FEIGHTNER, J.W. (1982) The clinical reasoning of randomly selected physicians in medical general practice, *Clinical and Investigative Medicine*, 5, pp. 49–55.

BORDAGE, G. (1994) Elaborated knowledge: a key to successful diagnostic thinking, *Academic Medicine*, 69, pp. 883–885.

BRAILOVSKY, C.A. & GRAND'MAISON, P. (2000) Using evidence to improve evaluation: a comprehensive psychometric assessment of a SP-Based OSCE licensing examination, *Advances in Health Sciences Education*, 5, pp. 207–219.

CHARLIN, B., BRAILOVSKY, C.A., BRAZEAU-LAMONTAGNE, L., SAMSON, L. & VAN DER VLEUTEN, C.P. (1998) Script questionnaires: their use for assessment of diagnostic knowledge in radiology, *Medical Teacher*, 20, pp. 567–571.

CHARLIN, B., TARDIF, J. & BOSHUIZEN, H.P.A. (2000a) Scripts and medical diagnostic knowledge: theory and applications for clinical reasoning instruction and research, *Academic Medicine*, 75, pp. 182–190.

CHARLIN, B., BRAILOVSKY, C.A., ROY, L. & VAN DER VLEUTEN, C.P. (2000b) The script concordance test: a tool to assess the reflective physician, *Teaching and Learning in Medicine*, 12, pp. 189–195.

CUSTERS, J.F.M., REGHER, G. & NORMAN, G.R. (1996) Mental representations of medical diagnostic knowledge: a review, *Academic Medicine*, 71, pp. S55–61.

ELSTEIN, A.S., SHULMAN, L.S. & SPRAFKA, S.A. (1978) *Medical Problem-solving: An Analysis of Clinical Reasoning* (Cambridge, MA, Harvard University Press).

FELTOVICH, P.J. (1983) Expertise: reorganizing and refining knowledge for use, *Professions Education Research Notes*, 4, pp.5–9.

FERLAND, J.J., BORDAGE, G. & LOISELLE, J-M. (1983) The effect of item translation on the psychometric status of a medical examination, *Annals of the Royal College of Physicians and Surgeons of Canada*, 16, pp. 641–646.

FRIEDMAN BEN-DAVID, M. (2000) The role of assessment in expanding professional horizons, *Medical Teacher*, 22, pp. 472–477.

MARSHALL, K.G., BRAILOVSKI, C.A. & GRAND'MAISON, P. (1995) French–English, English–French translation process of an objective structured clinical examination (OSCE) used for licensing family physicians in Quebec, *Teaching and Learning in Medicine*, 7, pp. 115–120.

McGAGHIE, W.C. (1993) Evaluating competence for professional practice, in: L. Curry & J.F. Wegin (Eds) *Educating Professionals. Responding to New Expectations for Competence and Accountability*, pp. 229–261 (San Francisco, Jossey-Bass).

NORCINI, J.J., SHEA, J.A. & DAY, S.C. (1990) The use of the aggregate scoring for a recertification examination, *Evaluation and the Health Profession*, 13, pp. 241–251.

NORMAN, G.R. (1985) Defining competence: A methodological review, in: V.R. Neufeld & G.R. Norman (Eds) *Assessing Clinical Competence*, Vol. 7, pp. 15–35 (New York, Springer).

NORMAN, G.R. (1985). Objective measurement of clinical performance, *Medical Education*, 19, pp. 43–47.

REGEHR, G. & NORMAN, G.R. (1996) Issues in cognitive psychology: implications for professional education, *Academic Medicine*, 71, pp. 988–1001.

SCHMIDT, H.G., NORMAN, G.R. & BOSHUIZEN, H.P.A. (1990) A cognitive perspective on medical expertise: theory and implications, *Academic Medicine*, 65, pp. 611–621.

VAN DER VLEUTEN, C.P.M. (1996) The assessment of professional competence: development, research and practical implications, *Advances in Health Sciences Education*, 1, pp. 41–67.

ZEITZ, C.M. (1997) Some concrete advantages of abstraction: how experts' representations facilitate reasoning, in: P.J. Feltovich, K.M. Ford & R.R. Hoffman (Eds) *Expertise in Context: Human and Machine*, pp. 43–65 (Menlo Park, CA, AAAI Press).