The Cognitive Validity of the Script Concordance Test: A Processing Time Study

Robert Gagnon
Bernard Charlin
Unit for Research and Development In Health Sciences Education
University of Montreal
Montreal, Canada

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

Monique St-Martin
Department of Medicine
University of Montreal
Montreal, Canada

Évelyne Sauvé
Unit for Research and Development in Health Sciences Education
University of Montreal
Montreal, Canada

Henny P. A. Boshuizen
Open University of the Netherlands
Educational Technology Expertise Center
Heerlen, The Netherlands

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

Background: *According to the theory on which the Script Concordance Test (SCT) is based, scripts contain expectations on features that are associated with each illness and about the range of values that are typical, atypical, or incompatible.*
Purpose: *To document the construct validity of the SCT, we investigated the theory prediction that once a script is activated, new incoming information (e.g., additional clinical features) is processed faster if it is typical for that script than if it is atypical. If it is incompatible, processing time falls in between.*
Methods: *We presented 2 groups of participants (30 fourth-year medical students and 30 full-time geriatricians) with 64 clinical vignettes (divided over 5 types of prevalent clinical presentations in geriatrics), each accompanied by a diagnostic hypothesis aimed to instantiate an appropriate script. Next, we presented a new finding, which could be typical, atypical, or incompatible given the hypothesis. Participants had to decide as quickly and accurately as possible whether the new finding increased, decreased, of did not affect the likelihood of the diagnostic hypothesis. We administered*

This research project was funded by a grant from the Medical Research Council of Canada/Association of Canadian Medical Colleges. We obtained approval for the study from the University of Montreal ethical review board.

Correspondence may be sent to: Bernard Charlin, URDESS, Faculté de médecine-direction, Université de Montréal, CP 6128, succursale centre-ville, Montréal, Québec H3C 3J7, Canada. E-mail: bernard.charlin@umontreal.ca

the test on a computer. The dependant variable was processing time. We analyzed data with a repeated measure 2 × 3 analysis of variance.

Results: *Typical information was processed faster than atypical and incompatible information (M = 10.6 sec vs. 19.2 sec and 16.4 sec, respectively; p < .001 for both). Incompatible information was processed faster than atypical information (16.4 sec vs. 19.2; p < .001). There was no significant difference between the groups of geriatricians and students.*

Conclusion: *It is possible to predict what kind of information will be processed faster depending of the typicality and compatibility of clinical data for given hypotheses. Results support SCT construct validity.*

Teaching and Learning in Medicine, *18*(1), 22–27

Feltovich and Barrows[1] and Schmidt, Norman, and Boshuizen[2] have described clinical competence development as a progressive construction of illness scripts. Charlin, Tardif, and Boshuizen[3] explained why script knowledge organization is particularly well fitted to explain clinical reasoning performance. According to theory, scripts are goal-directed knowledge structures adapted to perform tasks efficiently.[1,4]

The basic principle underpinning the script concept asserts that to give meaning to a new situation, people use prior knowledge that contains information about the characteristics and features of the situation and information about the relations that link those characteristics and features. In other words, incoming information activates a previously acquired network of relevant knowledge and experience—a script—that directs the selection, interpretation, and memorization of that new information.[5] In medicine, when a physician sees a patient, he or she perceives features—symptoms, signs, and details from the patients environment—that activate networks of knowledge that contains those features and their relationships to illnesses. Those networks of knowledge then provide context, and thus meaning, to the new situation.[3]

Smith[6] provided insights into how scripts are adapted to a diagnostic (categorization) goal. A script can be described as a set of attributes, each of which can be instantiated by values that have more or less probability of occurring. For each attribute, the value that has the greatest probability of occurrence is the default value. The script contains attributes (e.g., pain characteristics for a maxillary sinusitis script) for which different values are possible (no pain, dull sensation, infraorbital pain). In any given instance, only one of the values can fill the slot. Until the physician determines otherwise, the default value (in this case, infraorbital pain) is assumed to be present.[3]

Scripts are generic structures[6] that can represent any instance of an illness. Each medical encounter implies an instantiation process, that is, the finding of the actual values of the attributes observed in the patient. In a new situation, a set of relevant scripts is activated from the cues perceived, and the activity is then to find if one of the activated scripts adequately fits the clinical findings. This verification requires that values be assigned to the different attributes. If the physician can not adequately fit an activated script to the findings, he rejects it and begins to verify another one. Hence, according to the script concept, the fundamental aspect of understanding a situation is a hypothesis-testing activity.[3,6]

According to this theory, reasoning occurs as a series of qualitative judgements. Each of these judgements can be measured and compared to those of a reference panel of experienced practitioners. This is the basis of a method for assessing reasoning on ill-defined problems and in contexts of uncertainty called the Script Concordance Test (SCT).[7] For each item, an authentic and challenging clinical situation is described in a vignette accompanied by a diagnostic hypothesis aimed to instantiate an appropriate script. Next, a new finding is presented, which could be typical (the default value for the attribute), atypical, or incompatible given the hypothesis. Participants have to decide as quickly and accurately as possible whether the new finding increased, decreased, of did not affect the likelihood of the diagnostic hypothesis. Answers are captured by a Likert scale. An illustration of the test format is given in Figure 1.

Our goal in this study was to test a dimension of the construct validity of the test, that is, the reality of a difference in information-processing time as a function of its typicality and compatibility for a given illness knowledge activation. This study bears on results obtained by Custers, Boshuizen, and Schmidt[8] in a study aimed at providing evidence in favor of the development of script along the acquisition of professional experience. Custers et al.[8] found a difference on reaction time when typicality of information was manipulated. In this study, we wanted to verify if difference in reaction time could be predicted for the reasoning tasks on SCT items.

We hypothesized that if scripts contain expectations about clinical features of an illness, there will be a difference in information-processing time depending on the typicality or compatibility of that new data (a clinical feature) for the activated hypothesis. Information-processing time should be longer when the value of the feature is atypical in contrast with a feature that

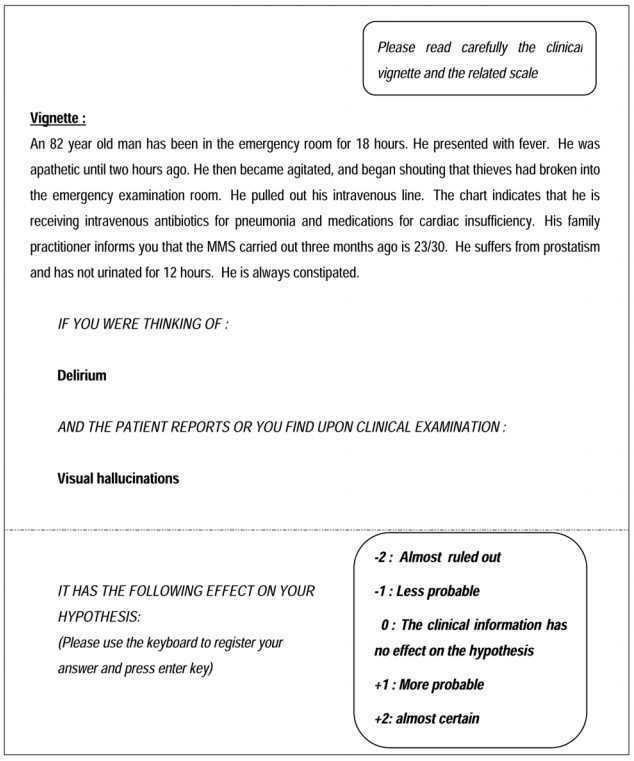

Figure 1. Format of item used for diagnostic knowledge assessment of dementia (computer screen).

is either typical or incompatible for that illness (atypical information will require more reasoning and consequently, extra processing time). We did the study using participants of two different levels of experience: students in their clinical clerkship and geriatricians. According to Schmidt et al.'s[2] conception of clinical reasoning development, scripts appear as soon as students are faced with time and pressure of clinical tasks and then are developed and refined with clinical experience.[3] Our second hypothesis was therefore that the effect should be present both for experts and for students and be stronger for experts than for students.

Method

Participants

For the students, 30 advanced medical students (4th year of studies in medicine) from the University of Montreal were asked to participate at the end of their 1-month rotation in geriatrics.

For the geriatricians, 30 full-time physicians with a practice in which the majority of patients belong to the geriatric population were recruited from Montreal and Sherbrooke Universities in Canada. They were identified and recruited with a snowball sampling methodology. All participants participated on a voluntary basis and were not paid. Among the geriatricians, 13 (43%) were specialized in geriatrics and 17 (57%) were family physicians having a significant part of their practice in geriatrics. Average number of years of practice was 13.6 years (*SD* = 7.3; range = 2–30).

Material

The material was based on five prevalent clinical presentations in geriatrics: falls, urinary incontinence, delirium, depression, and dementia. The acquisition of diagnostic competence for these common presentations is explicitly stated in students' educational objectives of the rotation in geriatrics. According to SCT construction methodology,[7] two experts (L. Roy & M. St-Martin) were asked (a) what are the most relevant hypotheses for that clinical presentation? and (b) what are the questions they ask, the physical exam they do, the kind of tests they ask in such situations? These experts were then asked what the values are they would consider as typical for an hypothesis (T), what are those they would consider as atypical (A) but still consider fitting, and what the values are that are incompatible with the hypothesis (I). When there was discrepancy in the classification between the two experts, the item was rejected. With this material, sets of items were constructed for each presentation. Each set began with a depiction of the clinical presentation within a short vignette. Each set had a balanced number of items of each category (T, A, and I). Across the whole test, the information that needed to be read in the different item categories (T, A, and I) was similar. For all situations, there were 12 items, with an exception for urinary incontinence that had 16 items. The material was prior tested with some clerkship students, residents, and experts for understanding and wording of items. Items within vignettes were presented in random order. A special set of items was built on a clinical presentation in another domain (obstetrics) to train participants on the task and the use of computer (training set).

Screen text appeared on computer in black on a white background (size of font = 18 point); no colors were used. The test lines were centered at the middle of the screen. A separate numeric keyboard was used to register the answers. A computer program was written to provide the text and to measure processing time (1 msec precision). Although we were interested by the time taken to answer and not by the answer itself, we asked participants to provide their best possible answer. Each item task appeared on screen in the format presented in Figure 1. Examples of item classification are provided in Table 1.

Procedure

We tested all participants individually. They were seated in front of a computer screen and given a short introduction. We then asked them to sign an informed

Table 1. *Example of Item Classification (for Items Related to the Clinical Scenario Presented in Figure 1)*

If You Were Thinking of	While You Find With the History or the Physical Examination	Classification of That Item
Delirium	Visual hallucinations	Typical
Delirium	Diminished attention	Atypical
Delirium	Normal level of consciousness	Incompatible

consent. Next, the experimenter started the session by saying "Read the vignette then press enter key to get the first question. Read the question thoroughly and answer it as best as you can." Participants were prepared with the training set and then passed through the whole 64 items. After completion, we debriefed them and gave them an opportunity to ask questions or to make remarks.

Measurement of processing time began every time the participants pressed the starting key (enter key) and was ended with pressure on the ending key (answer keys). Every time the starting key was pressed, the current item disappeared from the screen and the next item was presented. Thus, processing time on an item is the difference between the two keys pressing actions. Because the first question was attached and immediately following the text of the vignette, in that case, response time was a combination of time to read the vignette and time to answer the question. Thus, the time taken for the first answer was not used in the analysis. Processing time is made of reading time plus decision time plus the time needed to perform the physical response of pressing the key. The latter time is considered negligible. We informed all participants on the procedure in detail before the study and gave consent.

We measured the accuracy of answers on each category of typicality using the SCT scoring process described in the literature.[7,9] In this study, a panel of 10 out of 30 geriatricians was randomly chosen, and their answers on the 64 items were used to build the scoring scheme. The scoring process measures the concordance of examinees' answers with those of a panel of reference.

Statistical Analyses

We tested the research hypothesis by means of a repeated measure 2 × 3 analysis of variance (ANOVA) with clinical experience treated as a between-subject factor (two levels) and categories of typicality and acceptability treated as a within-subjects factor (three levels). The independent variable is the level of expertise, and the dependent variable is the mean response time to the three categories of typicality of information. This analysis allowed to test for the main effects of clinical experience and information type and to determine whether there was a significant interaction between the two factors. We set the significance level at .05. We checked assumptions underlying the analysis (e.g., normality, symmetry of the variance–covariance matrix), we used and the Greenhouse–Geisser correction in case of violation of the assumptions. Effect size (ES) is an index that measures the magnitude of a difference between two situations. It standardizes comparisons of the magnitude of group differences in different situations using standard deviations as yardsticks.[10]

Results

Figure 2 presents data on processing time for the 64 reflection tasks corresponding to items of the test. Mean response time and 95% confidence intervals are provided for the three types of information and for the two levels of experience. A main effect of type of information was observed, $F(2, 112) = 128.1$, $p < .0001$. No significant interaction was observed between levels of experience and type of information, $F(2, 112) = 0.17$, $p = .84$. There was no statistically significant effect of level of experience, $F(1, 56) = 0.91$; $p = .35$).

Because we observed no interaction, we combined data from both groups. We observed a significant effect of typicality of the information. For all participants, typical information was processed faster than atypical and incompatible information (M = 10.6 sec vs. 19.2

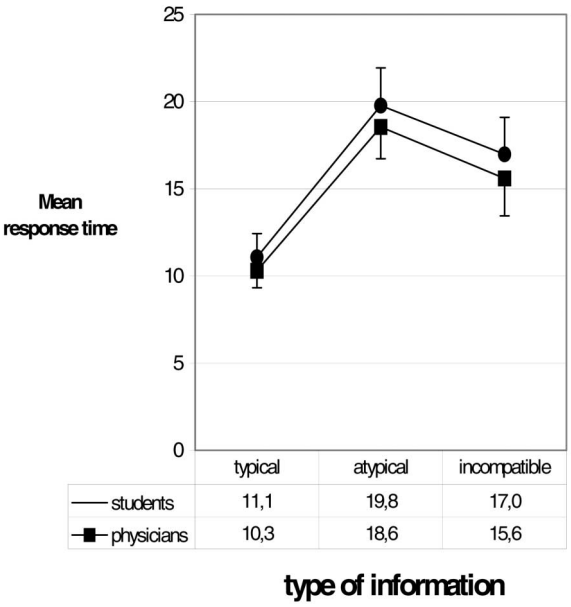

Figure 2. Overall reaction time in seconds (with 95% confidence interval).

sec and 16.4 sec, respectively; $p < .001$ for both). Incompatible information was processed faster than atypical information (16.4 sec vs. 19.2, respectively; $p < .001$). Differences of processing time, expressed in terms of ES (measurement unit is the pooled standard deviation) was 1.7 for typical versus atypical, 1.3 between typical and incompatible, and 0.4 between atypical and incompatible.

Accuracy of answers varied for each category of typicality: typical (76.3%; *SD* = 9.2), atypical (67.5%; *SD* = 6.5), and incompatible (67.9%; *SD* = 11.8). Statistically significant differences were observed between answers to typical items versus atypical and incompatible but not between the two latter categories. These patterns of responses were similar for both groups of respondents.

Discussion

The aim of the study was to verify if difference in reaction time could be predicted for reasoning tasks presented in the format of SCT items. ANOVA analyses of the data showed a clear picture for the processing of clinical information. Information was processed much more rapidly when it was made of features that are typical for the situation rather than atypical or incompatible. The difference in processing time was large: 1.7 and 1.3 *SD*s, respectively. It is noticeable that incompatible information was processed faster than atypical information (difference of 0.4 *SD*s). These results confirm the reality of a difference in information processing as a function of its typicality and compatibility when the knowledge for a given illness is activated.

Accuracy of answers was also significantly higher for typical items than for atypical and incompatible situations. This pattern was similar in both groups. This provides further support to the theory on which SCT is built. We did not find accuracy difference between the two levels of experience. This this may be a consequence of items deliberately built to detect differences in reaction time rather than to discriminate along levels of experience.

Our second hypothesis concerned processing-time differences for the two groups. In his study, Custers et al.[8] found that clinical experience was reflected by shorter mean processing time for family physicians. In our study, the analysis of processing time, whatever the nature of information, showed no statistically significant difference between the groups (14.8 sec for geriatricians vs. 16.0 sec for students). This absence of difference may be explained by two reasons. First, students had several months of clinical exposure and tested knowledge was part of (or was taken from) the educational objectives of students' rotation in geriatrics. It is possible that at the end of rotation, students had developed scripts efficient enough to well address the tasks. Similarity of accuracy of answers in both groups is in line with this last possibility. The second reason concerns the nature of the cognitive task. Assessing whether new information is consistent with an activated illness script (this is what the reaction times in the Custers et al. study[8] actually represent) is quite different from gauging the consequence of new information for the appropriateness of this script. This task requires additional processing time. We actually found that reaction time in our study was much longer, generally in the 10 to 20 sec range, as opposed to 2 to 6 sec in the Custers et al. study.[8] This extra task component may have overshadowed differences between the advanced students and the geriatricians.

Elstein et al.[11] 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, judgements, and choices to those of the experienced group. The script concordance approach is designed according to these principles. As such, it represents a shift in the strategy of clinical reasoning assessment. In the last decades, the strategy has often been realized by a transposition of clinical encounters on paper, on computer, or with simulated patients (Objective Structural Clinical Exam). Thus, the script concordance approach, instead of trying to simulate the encounter and assess reasoning outcomes, attempts to place examinees in a specific context and probes cognition processes.

The aim of this study was to verify that the test format really allows probing cognition processes. This verification was done by checking assumptions taken from theory. Results show that it is possible to predict what kind of information will be processed faster depending on the typicality and compatibility of clinical data for given hypotheses, thus providing arguments on the construct validity of the test.

SCT is a theory-based test. This study strengthens its foundations. It is part of a research agenda made of issues such as discrimination according to level of experience,[7] validity of the scoring process,[12] predictive validity,[13] stability of the test across two different linguistic and learning environments,[14] measurement of perception and interpretation skills,[15] assessment in ethics,[16] the meaning of the variability of panel members' answers as a way to detect levels of clinical experience,[17] or the number of members needed on panels of reference to obtain reliable scores.[9] Future studies will be directed toward the development of tools administered online for assessment of analytic and nonanalytic skills in visual domains; toward use of the test as a teaching strategy, notably for detection and remediation of students/residents with reasoning difficulties; and toward other studies on psychometric issues such as standard setting, that is, the definition of a

reproducible and defendable procedure to establish cutting scores in a population of examinees.

References

1. Feltovich PJ, Barrows, HS. Issues of generality in medical problem solving. In HG Schmidt (Ed.), *Tutorials in problem-based learning: A new direction in teaching the health professions* (pp. 128-142). Assen, Holland: Van Gorcum, 1984.
2. Schmidt HG, Norman GR, Boshuizen HP. A cognitive perspective on medical expertise: Theory and implication. *Academic Medicine* 1990;65:611–21.
3. 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.
4. Nelson K. *Event knowledge: Structure and function in development.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc., 1986.
5. Schallert DL. The significance of knowledge: A synthesis of research related to schema theory. In W Otto, S White (Ed.), *Reading expository prose* (pp. 13–48). New York: Academic, 1982.
6. Smith EE. Concepts and induction. In MI Posner (Ed.), *Foundations of cognitive science*. Cambridge, MA: MIT Press, 1989.
7. Charlin B, van der Vleuten C. Standardized assessment of reasoning in contexts of uncertainty: The script concordance approach. *Evaluation & The Health Profession* 2004;27(3):304–19.
8. Custers EJ, Boshuizen HP, Schmidt HG. The influence of medical expertise, case typicality, and illness script component on case processing and disease probability estimates. *Memory & Cognition* 1996;24:384–99.
9. Gagnon R, Charlin B, Coletti M, Sauvé E, van der Vleuten C. Assessment in the context of uncertainty: How many members are needed on the panel of reference of a script concordance test? *Medical Education* 2005;39:284–91.
10. Cohen J. *Statistical power analysis for the behavioral sciences.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc., 1988.
11. Elstein AS, Shulman LS, Sprafka, SA. Medical problem solving, a ten-year retrospective. *Evaluation & The Health Profession* 1990;13:5–36.
12. Charlin B, Desaulniers M, Gagnon R, Blouin D, van der Vleuten C. Comparison of an aggregate scoring method with a consensus scoring method in a measure of clinical reasoning capacity. *Teaching and Learning in Medicine* 2002;14:150–6.
13. Brailovsky C, Charlin B, Beausoleil S, Coté S, Van der Vleuten C. Measurement of clinical reflective capacity early in training as a predictor of clinical reasoning performance at the end of residency: An exploratory study on the Script Concordance Test. *Medical Education* 2001;35:430–6.
14. Sibert L, Charlin B, Corcos J, Gagnon R, Grise P, Van der Vleuten C. Stability of clinical reasoning assessment results with the script concordance test across two different linguistic, cultural and learning environments. *Medical Teacher* 2002;24:537–42.
15. Brazeau-Lamontagne L, Charlin B, Gagnon R, Samson L, van der Vleuten C. Measurement of perception and interpretation skills along radiology training: Utility of the script concordance approach. *Medical Teacher* 2004;26:326–32.
16. Llorca G. Évaluation de résolution de problèmes mal définis en éthique clinique: Variation des scores selon les méthodes de correction et selon les caractéristiques des jurys [Assessment of ill-defined problems in clinical ethics: Variation of scores depending of scoring methods and of characteristics of the jury]. *Pédagogie Médicale* 2003;4:80–8.
17. Charlin B, Gagnon R, Pelletier J, et al. Assessment in context of uncertainty: The effect of variability within the panel of reference. *Medical Education.* In press.

Received 22 November 2004
Final revision received 27 July 2005