

Current written tools of assessment are mostly measuring the capacity to solve well-defined problems by the application of rules and principles, while the essence of expertise in the professions lies in the capacity to solve ill-defined problems, that is, reasoning in contexts of uncertainty. The purpose of this study is to describe an approach that allows assessing ill-defined problems and to present and discuss research findings related to this approach. The tool has been used up to now mainly in medicine, however it can be applied in all health professions. The approach is based on three principles: (a) examinees are faced with a challenging authentic situation in which several options are relevant; (b) the response format is a Likert-type scale that reflects the way information is processed in problem-solving situations, according to the script theory; and (c) scoring is based on the aggregate scoring method to take into account the variability of reasoning processes among experts. Research findings suggest that the approach permits one to reliably discriminate examinees across their level of experience, and so in very different domains. It makes it possible to measure skills or domains that were up to now difficult to measure.

Keywords: *assessment of reasoning; uncertainty; health professions; medicine; ill-defined problems; script concordance approach*

STANDARDIZED ASSESSMENT OF REASONING IN CONTEXTS OF UNCERTAINTY

The Script Concordance Approach

BERNARD CHARLIN

University of Montreal

CEES VAN DER VLEUTEN

Maastricht University

AUTHORS' NOTE: The development of the approach has been funded by grants from the Medical Research Council of Canada/Association of Canadian Medical Colleges, the Medical Council of Canada and the Royal College of Physicians and Surgeons of Canada. All studies obtained ethical approval.

Reasoning in the professions is much more than simple application of knowledge, rules, and principles. Schön (1983) made a distinction between the kinds of problems professionals encounter in their practice. For some problems, all data necessary to solve them are present, the goals to reach are clear, and there are known solutions to reach the goals. These problems, named well defined, can be solved using knowledge and skills that Schön considered as belonging to the domain of technical rationality. A significant part of professional reasoning competence rests on the capacity of applying well-known solutions to well-defined problems. Written tests of excellent reliability and validity such as multiple-choice questions (MCQ) tests assess successfully the “technical rationality” part of professional reasoning. However, in a clinical encounter, not all the data necessary to solve a problem are available. These data must be gathered to formulate the problem and then solve it. Furthermore, problems can be confusing and contradictory and are characterized by imperfect, inconsistent, or even inaccurate information. Problems are often ill defined and characterized by uncertainty (Fox, 2000). The capacity to reason in contexts of uncertainty and to solve ill-defined problems is the hallmark of professional competence (Johnson, 1988), and the knowledge needed to successfully reason in these contexts is called professional knowledge (Schön, 1983). Some evaluation tools, oral exams for example, can assess that kind of knowledge, however these tools have their limitations such as difficulty of standardization, of scoring objectivity, or of practicability for large groups of examinees.

A difficulty with assessment on ill-defined problems is that, as shown in medicine, in similar situations professionals do not collect exactly the same data and do not follow the same paths of thought (Grant & Marsden, 1988). They also show substantial variation in performance on any particular real or simulated case (Barrows, Feightner, Neufeld, & Norman, 1978; Elstein, Shulman, & Sprafka, 1978). This variability must be taken into account in the assessment of ill-defined problems. The commonly used MCQ test format requires a unique right solution to problems, a right conclusive answer to give to data specified in the problem. When examination jury members are unable to agree on a unique right answer on an item, the item is often removed from the examination. This impairs assessment of ill-defined problems, the hallmark of authentic problem solving in the professions.

Most current methods of professional competence assessment, either performance-based methods (Dauphinee, 1995) (e.g., Objective Structured Clinical Exams) or methods assessing the solutions found to well-defined problems (e.g., MCQs), are measures of behavior. At a time when cognitive psychology has become the major conceptual framework in educating to the professions (Irby, 1997), it is necessary to add to these methods a way to assess reasoning cognition. We should also measure its process instead of its outcome, and this measurement should be based on theory. The adaptation of cognitive psychology script theory (Charlin, Tardif, & Boshuizen, 2000; Schmidt, Norman, & Boshuizen, 1990) to the characteristics of reasoning in the professions provides a promising way to build a theory-based assessment tool.

A script is a goal-directed knowledge structure adapted to perform tasks efficiently (Nelson, 1986). Scripts begin to appear when students are confronted with real professional tasks (Schmidt, Norman, et al., 1990). They are then developed and refined during one's professional life (Schmidt, Norman, et al., 1990). The theory implies that to give meaning and to act adequately to a situation, professionals activate scripts relevant to the situation. These structures are used to actively process information to confirm or eliminate hypotheses, or management options (Charlin, Tardif, et al., 2000). According to this theory reasoning is made with a series of qualitative judgments. Each of these judgments can be measured and compared to those of a reference panel of experienced practitioners. This provides a method of assessment of reasoning on ill-defined problems and in contexts of uncertainty (Charlin, Roy, Brailovsky, & van der Vleuten, 2000). It is named the script concordance approach.

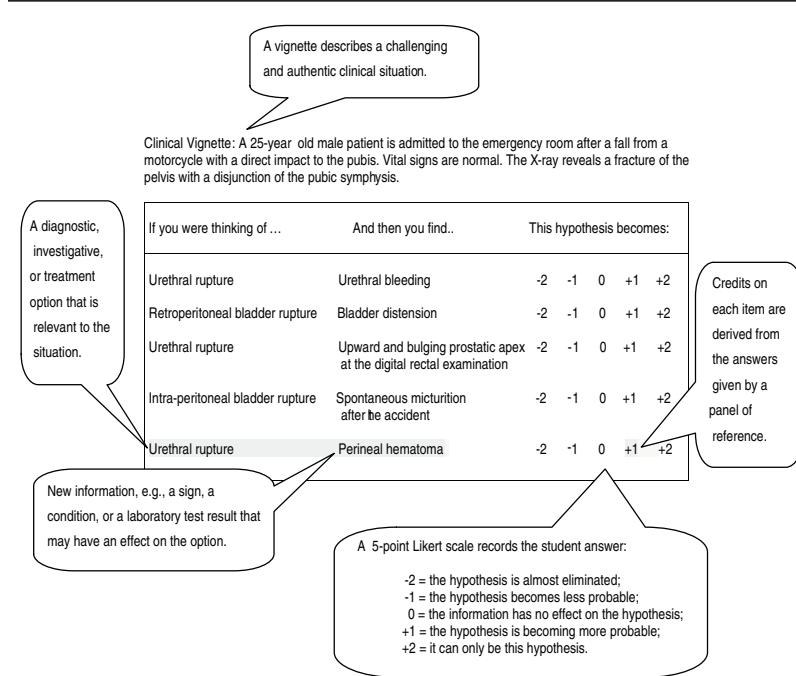
The approach allows making different kinds of tests, depending on the field one wishes to assess. Up to now the approach has been used mainly in medicine, however it is being used in pharmacy (Khonputsa, Besinque, Fisher, & Gong, 2004), rehabilitative therapy (Cohen, 2003), and midwifery (A. Demeester, personal communication, November 2003). The goal of this article is to share with colleagues of other health professions results provided by a series of studies addressing questions concerning the validity, reliability, and usefulness of the test in the educational settings.

THE SCRIPT CONCORDANCE APPROACH

To assess ill-defined problems, the approach rests on three principles, each of them concerning one of the three components any test has: the task required of examinees, the way examinees' answers are recorded, and the way examinees' performances are transformed to a score (Norman, Swanson, & Case, 1996). The task is challenging, even for an expert. It represents an authentic clinical situation and is described in a vignette. It is challenging either because the vignette does not contain all the data needed to provide a solution (for a test on diagnosis or management for instance) or because several attitudes are defensible (for a test in ethics for instance). Several options (diagnosis, management, or attitude) are relevant. Items are questions experts ask themselves to progress toward a solution. The response format is in accordance with what is known from clinical reasoning processes (Barrows et al., 1978; Elstein et al., 1978; Grant & Marsden, 1988). A Likert-type scale records the judgments that are constantly made within this process (Charlin, Tardif, et al., 2000). The scoring method takes into account variation of answers among jury members. It is an adaptation of the aggregate scoring method (Norcini, Shea, & Day, 1990; Norman, 1985). The method to build tools according to the script concordance approach is described in detail elsewhere (Charlin, Roy, et al., 2000).

The item format differs with the objective of assessment (e.g., measurement of competence for diagnosis, investigation, or treatment, measurement of attitudes). The clinical situation is described in a vignette, and for a given vignette, items are regrouped by formats (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, for example, a sign, a condition, an imaging study, or a laboratory test result that may have an effect on the diagnostic hypothesis, the investigative action, or the treatment option. The third part is a 5-point Likert-type scale that records the student answer. An illustration of the format is given in Figure 1.

The aggregate scoring method (Norcini et al., 1990; Norman, 1985) used with the test reflects the variability experts demonstrate in

**Figure 1: The Test Format**

their reasoning processes. Credits on each item are derived from the answers given by a panel of reference. The credit for each answer is the number of panel members that have provided that answer, divided by the modal value for the item. For example, if on an item six panel members (out of 10) have chosen response +1, this choice receives 1 point (6/6). If three experts chose response +2, this choice receives 0.5 (3/6), and if one expert chose response 0, this choice receives 0.16 point (1/6). The total score for the test is the sum of credits obtained on all items. This score is then divided by the number of items and multiplied by 100 to get a percentage score.

RESEARCH FINDINGS

The approach has been tested in several studies to address issues such as construct validity, reliability, and feasibility of the approach;

validity of its scoring process; and its applicability in different clinical contexts. In this section we present findings that are crucial in the process of a test development (e.g., discriminative power across levels of experience) and findings that indicate a potential for assessment in domains that were up to now difficult to assess with standardized tools (e.g., perception and interpretation skills in imaging specialties, or attitudes in ethics).

DISCRIMINATION ACCORDING TO LEVEL OF EXPERIENCE

A tool aimed at measuring clinical reasoning competence should overcome the intermediate effect (Schmidt, Boshuizen, & Hobus, 1988), a limitation found with some test formats based on written simulations of clinical problem solving. It is a puzzling fact that experienced clinicians judged competent by peers, often perform little better, and sometimes worse than clinicians with intermediate levels of experience (end-of-training residents, for instance).

Early studies on the Script Concordance Test (SCT) (Charlin, Brailovsky, Brazeau-Lamontagne, Samson, & Leduc, 1998; Charlin, Brailovsky, Leduc, & Blouin, 1998) were undertaken to verify the discriminant validity of the test. Results showed an increase in the mean scores of individuals with differing levels of clinical expertise (students, residents, and staff members), the less experienced getting the lower results. This finding was consistently found in subsequent SCT studies. It supports the construct validity of the instrument, indicating that the SCT measures a dimension for which, as one should expect, experienced physicians get better scores than less experienced participants.

It is interesting to note that all studies were done without any specific revision of subject matter. Examinees, whatever their level of experience, took their test with the knowledge they had at that moment, exactly as they do in their clinical work when they have an encounter with a patient. The test was nevertheless able to discriminate participants according to their level of experience.

VALIDITY OF THE SCORING PROCESS

An aggregate scoring method has been proposed by Norman (1985) and by Norcini et al. (1990) in the late 1980s. The method was

conceived to address the important issue of variability of experts' reasoning paths when they reason on ill-defined problems. Nevertheless, despite the importance of this issue, the method had since seldom been used or studied in the literature. An SCT study was conducted in gynecology (Charlin, Desaulniers, Gagnon, Blouin, & van der Vleuten, 2002) to determine if this method is superior to conventional methods. A test of 45 items was built to test on three different clinical situations (first trimester bleeding, abnormal uterine bleeding, and request for a contraceptive method). At the end of their rotation in gynecology-obstetrics, 150 students passed the test. The seven staff members of the department of the school were asked to be the reference panel. Seven other experts (tested experts) from another school volunteered to participate in the study as examinees.

Answer keys were built with two methods. Members of the panel of reference were first asked to complete the test individually. Their answers were used to build an answer key through an aggregate scoring method. One year later, the same persons were asked to meet and provide the "right answer" by consensus for each item. This scoring method was called the consensus method. Panel experts' answers on items varied substantially in the two contexts. Fifty-nine percent of the time the answer differed when they were asked to complete the test individually and when they were asked to provide, in a group meeting, the right answer to require from students. An explanation for this might be that the context for the tasks differs radically in the two conditions. When an expert is alone (as it is usually in practice), only the data provided by the case are used, while within a group the context is significantly modified by interactions with the other experts. This implies that for answer key construction, experts should be placed in the same conditions as future examinees.

Scores obtained with the two methods by students and tested experts were then compared. The aggregate method provided higher scores to tested experts and allowed a better discrimination of scores among examinees. The mean difference between students and tested experts was statistically significant. With the consensus method, the mean difference between groups was not statistically significant. The aggregate method was therefore superior to the consensus method with respect to the provision of scores in ill-defined problem assessment and in contexts of uncertainty.

PREDICTIVE VALIDITY

A study was carried out to verify whether scores obtained on a SCT taken at the end of clerkship predict those obtained on tests of clinical reasoning 2 years later at the end of residency (Brailovsky, Charlin, Beausoleil, Côté, & van der Vleuten (2001). In family medicine residency in Canada, the certification exam comprises three tests of clinical competence. The first, named Short Answer Management Problems (SAMP) is composed of 42 clinical vignettes, each of them 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. The second, named Simulated Office Orals (SOO) is composed of five simulations of clinical encounters. Each 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 pretested objective marking scheme. The test assesses the candidate's abilities to manage complex biopsychosocial problems, with an emphasis on the patient-doctor relationship. The third is an Objective Structured Clinical Examination (OSCE) made of a series of 13 stations in which distinct clinical skills (e.g., history taking, physical exams, or technical skills) are assessed. By virtue of test construction, the SOO and the SAMP (at least in part) measure output (SAMP), and output and process (SOO) of clinical reasoning while the OSCE measures much more hands-on clinical skills than clinical reasoning.

A cohort of 24 students from a medical school was followed up to the end of their residency in family medicine. The authors assumed that the adaptation of knowledge organization to clinical tasks, as indicated by SCT scores would predict part of the performance on the measures of clinical reasoning (SAMP and SOO) but would be less predictive of the OSCE. (All three tests were taken 2 years after the administration of the SCT.) Data found in the study were consistent with the hypothesis. Pearson correlation coefficients were statistically significant when comparing scores on the SCT with those of the SAMP and the SOO, respectively. When the correlation was made with the OSCE, however, there was no statistical significance. One interpretation is that, if a candidate has shown good organization of his or her clinical knowledge at an early moment in training, it can be

expected that he or she will show good organization at subsequent measurements, even if the newly tested domain differs from the first one.

STABILITY OF THE TEST ACROSS TWO DIFFERENT LINGUISTIC AND LEARNING ENVIRONMENTS

An 80-item test was built, based on the major educational objectives of Canadian and French urology training programs and from clinical situations representative of urology practice (Sibert et al., 2002). Care was taken to ensure the content validity of the test and the quality of translation of the test in the two languages. The 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 university, 8 from the Canadian university). Two groups of certified urologists made the reference panels: 10 French urologists and 12 Canadian urologists. Each urologist had a minimum of 5 years clinical experience, and all of the Canadian participants were English speaking.

The two groups of urologists were employed as the reference panel for the construction of answer keys. Reliability was assessed via Cronbach's alpha coefficient. Scores between groups were compared by analysis of variance. Reliability coefficients of the 80-item test were .794 for the French participants and .795 for the Canadian participants. Scores increased with clinical experience in urology in the two sites, and candidates obtained higher scores when the answer key provided by the experts of the same country was employed. These data provide support for the construct validity of the tool across different learning environments.

MEASUREMENT OF PERCEPTION AND INTERPRETATION SKILLS

A study was done in radiology (Brazeau-Lamontagne, Charlin, Gagnon, Samson, & van der Vleuten, *in press*) to verify if, with the script concordance approach, it would be possible to reliably measure skills that are traditionally difficult to measure: perception and interpretation skills in film reading. A perception test (PT) and an interpretation test (IT) were built according to the script concordance approach. Both tests were drawn from the same radiology domain:

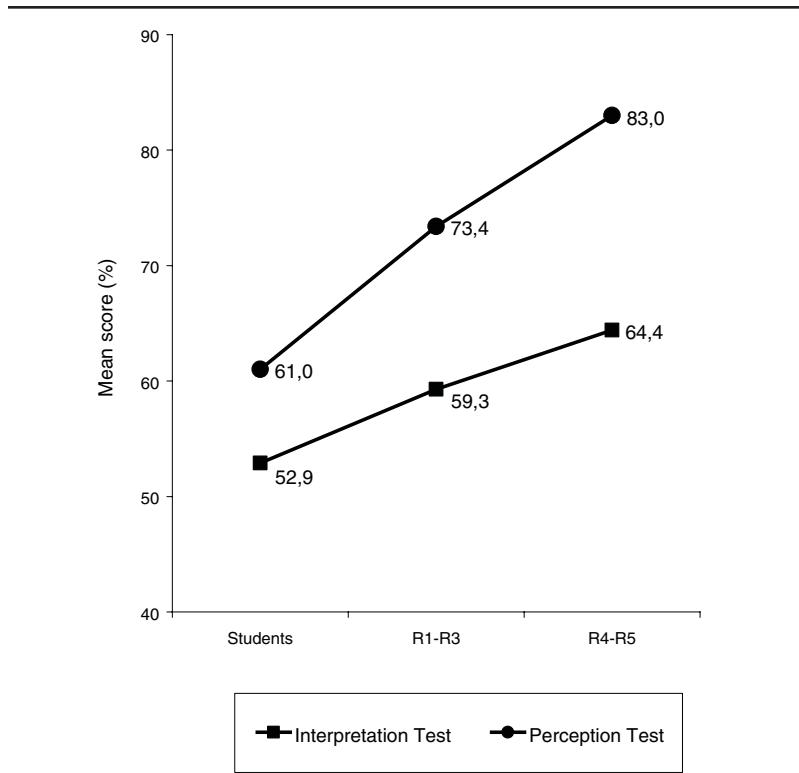


Figure 2: Perception and Interpretation Scores According to Level of Training

plain chest X-ray reading. Groups representing three levels of radiological training were tested: clerkship students (20), junior residents (R1 to R3: 20), and senior residents (R4 and R5: 20). The PT was composed of 38 items based on four radiological problems represented by four sets of chest X-rays. The IT was composed of 145 items (regrouped for analyses at the level of 29 signs) related to the same four problems presented on four other sets of chest X-rays. Answer keys were constructed with the aggregate scoring method on answers built with answers provided by 11 board-certified radiologists currently reading chest X-rays.

The alpha coefficients on the 38 items of the PT were .79 and .81 for the IT. The differences between mean scores obtained by the three groups with the PT and the IT were analyzed with an ANOVA. In both

tests all differences between groups were highly significant ($< .001$) while all of the means showed a constant progression from the lower level of training (students) to the higher level (R4 and R5). Perception scores tended to increase more rapidly than interpretation scores (see Figure 2). The two tests demonstrated large effect sizes (2.2 for the PT and 1.6 for the IT) in the discrimination of lower versus higher levels of expertise.

ASSESSMENT IN ETHICS

Llorca (2003) adapted the script concordance approach to assess ethical opinions and attitudes in the professions by constructing a test to assess attitudes toward a difficult therapeutic decision. A vignette describing the clinical problem was paired with a presentation of six attitudes that could be legitimately defended. Examinees had to provide, on a Likert-type scale, their opinion about the relevance of each attitude. The test was passed by 61 examinees (50 students and 11 residents in rheumatology). Four panels were made with 139 persons: the ethics committee of a university hospital (15), professors of different specialties with a specific interest in ethics (15), professors of therapeutics (44), and family physicians (65). The four different panels of reference reflected different ethical perspectives. Opinion tendencies among them were studied, and answer keys constructed by these four panels were used to score the performance of residents and students.

Three of the four panels demonstrated similar opinion tendencies, although the family physician panel had a more technical perspective. The tendency of residents' opinions was relatively close to those of the family physician panel, while those of students differed significantly from those of all professional panels. When opinions were transformed into scores, it appeared that residents' scores were significantly higher than those of students, whichever the panel of reference. These findings indicate that the script concordance approach may allow testing in a domain that has hardly been assessed up to now, such as difficult therapeutic situations that implicate ethical judgment. Items become judgments on the relevance of several attitudes toward the ethical problem. The test then becomes an opinion survey in which the aggregate scoring method assigns a score to participants reflecting how close they are from the opinion of the different panels of experts.

RELIABILITY, FEASIBILITY, AND EDUCATIONAL QUALITIES OF THE TEST

Across studies several qualities of the test have been identified. A test is often considered to be sufficiently reliable when its Cronbach's alpha coefficient reaches a value of .80. In a series of studies (Brailovsky et al., 2001; Sibert et al., 2002; Charlin, Brailovsky, Brazeau-Lamontagne, et al., 1998;) values ranged from .79 to .82 with relatively small numbers of items (29 to 80). Experience shows that an 80-item test can be passed in 1 hour or less. This compares very favorably with the time required by other examination formats to reach the .80 values.

A good assessment tool should be relatively easy to build, easy to administer, and easy to score. At first glance it may appear simple and rapid to construct script concordance tests because only a relatively few experts for any given domain need to (a) describe clinical situations that are characteristic of the domain to assess and (b) specify the questions they would ask and the action they would take in the situation to arrive at a diagnosis (or decide on the adequate management of the patient). Items are then built with this material. In reality, however, writing items for the SCT requires some familiarity with the tool and acquisition of some skills. For instance, (a) the situation described must be complex enough to be challenging for the level of training that has to be assessed; (b) there must not be enough data to "solve" it, even for an expert; and (c) data provided in the items, however, must be data that an expert may seek to progress toward a solution. Still, construction of the SCT is probably no more difficult than the construction of a high-quality multiple-choice test.

Validation tasks and tasks related to elaboration of answer keys must also be well accepted by experts. These phases need not be inordinately time-consuming, however, and physicians that constitute panels appear to enjoy completing a test that is close to real clinical reasoning. In addition, the sometimes-lengthy discussions required to arrive at consensus answers in other testing formats is not required. Test administration can be paper or computer based, without particular difficulty. The scoring system is simple.

In addition, SCTs can be used as a viable educational tool as witnessed by Labelle et al.'s description of a method for use in continuing medical education sessions (Labelle, Beaulieu, Thivierge, Paquette, & Choquette, 2001). Here, a test is built to address specific

educational objectives. In the session participants answer the test individually, then have small group discussions with the goal to produce answers by consensus. These answers are then presented on transparencies and compared to the one made by the panel of reference. Exchanges between participants and the expert leading the session are based on this comparison. If there are no differences on an item, then there is no need for further discussion on this subject, however if there are discrepancies then a discussion ensues. Experience proves that experts have to provide to participants strong arguments to convince them to adopt their way of dealing with the situation, which ensures lively training sessions.

DISCUSSION

In cognitive research on medical expertise, there has been a transition from the search for a generic problem-solving skill toward a focus on memory organization, knowledge use, and problem representation and how these change with experience. More work on the applications of these approaches is definitely needed, however. From this perspective, Elstein, Shulman, and Sprafka (1990) 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 script concordance approach is designed to accomplish this task. As such, it represents a shift in the strategy of clinical reasoning assessment. In the past decade, the strategy has often been realized by mimicking reality as much as possible by a transposition of clinical encounters either on paper or on computer. The script concordance approach thus, instead of trying to simulate the encounter and assess reasoning outcomes, attempts to place examinees in a specific context and probes cognition processes.

The studies presented in this article illustrate the potential of this new assessment approach. The approach appears to be able to discriminate examinees according to their level of experience in very different domains including family medicine and surgical, medical, and imaging specialties. It also seems able to measure skills or domains that up to now have been difficult to measure. This offers a whole range of applications, such as assessment of intraoperative decision-

making skills and assessment of therapeutic decisions in contexts in which evidence-based medicine cannot be applied. These two applications are currently under study.

Nevertheless many questions remain. Some questions concern the panels of experts, such as the number of experts needed to provide stable scores. Other questions that need to be answered include:

- What rules should govern the choice of panel members?
- Is the current 5-point Likert-type scale the optimal format?
- If all members agree on the answer, is an MCQ indicated rather than an SCT?
- If there is total disparity, is this noise rather than a signal?
- Are the items that possess a middle range of variance among panel members' answers the ones that allow detection of experience among examinees?
- How can a standard-setting procedure be applied to this type of test?

CONCLUSION

The script concordance approach is designed to measure reasoning on authentic problematic tasks. It produces tests that are standardized and objective with respect to scoring. Although considerably more research is needed, data exist suggesting its utility as a strategy for investigating the process of clinical reasoning within the health professions, perhaps in conjunction with other tools such as MCQs.

REFERENCES

- Barrows, H. S., Feightner, J. W., Neufeld, V. R., & Norman, G. R. (1978). *Analysis of the clinical methods of medical students and physicians* (Final Report, Ontario Department of Health, Grants ODH-PR-273 and ODH-DM-226). Hamilton, Canada: McMaster University.
- Brailovsky, C., Charlin, B., Beausoleil, S., Côté, S., & van der Vleuten, C. (2001). 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*, 35, 430-436.
- Brazeau-Lamontagne, L., Charlin, B., Gagnon, R., Samson, L., & van der Vleuten, C. (in press). Measurement of perception and interpretation skills along radiology training: utility of the script concordance approach. *Medical Teacher*.
- Charlin, B., Brailovsky, C. A., Brazeau-Lamontagne, L., Samson, L., & Leduc, C. (1998). Script questionnaires: Their use for assessment of diagnostic knowledge in radiology. *Medical Teacher*, 20, 567-571.

- Charlin, B., Brailovsky, C. A., Leduc, C., & Blouin, D. (1998). The Diagnostic Script Questionnaire: A new tool to assess a specific dimension of clinical competence. *Advances in Health Sciences Education*, 3, 51-58.
- Charlin, B., Desaulniers, M., Gagnon, R., Blouin, D., & van der Vleuten, C. (2002). Comparison of an aggregate scoring method with a consensus scoring method in a measure of clinical reasoning capacity. *Teaching and Learning in Medicine*, 14, 150-156.
- Charlin, B., Roy, L., Brailovsky, C. A., & van der Vleuten, C. P. M. (2000). The Script Concordance Test: A tool to assess the reflective clinician. *Teaching and Learning in Medical Education*, 12, 189-195.
- Charlin, B., Tardif, J., & Boshuizen, H. P. A. (2000). Scripts and medical diagnostic knowledge: Theory and applications for clinical reasoning instruction and research. *Academic Medicine*, 75, 182-190.
- Cohen, L. (2003). *The development and validation of the Seating and Mobility Script Concordance Test (SMSCT)*. Unpublished doctoral dissertation, University of Pittsburgh, PA.
- Dauphinee, W. D. (1995). Assessing clinical performance: Where do we stand and what might we expect? *Journal of the American Medical Association*, 274, 741-743.
- Elstein, A. S., Shulman, L. S., & Sprafka, S. A. (1978). *Medical problem solving: An analysis of clinical reasoning*. Cambridge, MA: Harvard University Press.
- Elstein, A. S., Shulman, L. S., & Sprafka, S. A. (1990). Medical problem solving, a ten-year retrospective. *Evaluation & the Health Professions*, 13, 5-36.
- Fox, R. C. (2000). Medical uncertainty revisited. In G. L. Albrecht, R. Fitzpatrick, & S. C. Scrimshaw (Eds.), *Handbook of social studies in health and medicine* (pp. 409-425). London: Sage.
- Grant, J., & Marsden, P. (1988). Primary knowledge, medical education and consultant expertise. *Medical Education*, 22, 173-179.
- Irby, D. (1997). Editorial. *Academic Medicine*, 72, 116.
- Johnson, E. J. (1988). Expertise and decision under uncertainty: Performance and process. In M. T. H. Chi, R. Glaser, & M. J. Farr (Eds.), *The nature of expertise* (pp. 209-228). Hillsdale, NJ: Lawrence Erlbaum.
- Khonputsa, P., Besinque, K., Fisher, D., & Gong, W. C. (2004). *The Pharmacy Script Concordance Test: A pilot study assessing clinical pharmacy competence in Thailand*. Manuscript submitted for publication.
- Labelle, M., Beaulieu, M., Thivierge, R. L., Paquette, D., & Choquette, D. (2001, January 27-30). *Best practice in CME: An innovative problem-based learning model based on the use of the Script Concordance Test*. Poster presented at the Alliance for Continuing Medical Education's 25th Annual Conference. San Francisco.
- Llorca, G. (2003). Évaluation de résolution de problèmes mal définis en éthique clinique: variation des scores selon les méthodes de correction et selon les caractéristiques des jurys [Ill-defined problem assessment in clinical ethics: Score variation according to scoring method and jury characteristics]. *Pédagogie Médicale*, 4, 80-88.
- Nelson, K. (1986). *Event knowledge: Structure and function in development*. Hillsdale, NJ: Lawrence Erlbaum.
- Norcini, J. J., Shea, J. A., & Day, S. C. (1990). The use of the aggregate scoring for a recertification examination. *Evaluation and the Health Professions*, 13, 241-251.
- Norman, G., Swanson, D. B., & Case, S. M. (1996). Conceptual and methodological issues in studies comparing assessment formats. *Teaching and Learning in Medicine*, 8, 208-216.
- Norman, G. R. (1985). Objective measurement of clinical performance. *Medical Education*, 19, 43-47.
- Schmidt, H. G., Boshuizen, H. P. A., & Hobus, P. P. M. (1988). Transitory stages in the development of medical expertise: The "intermediate effect" in clinical case representation studies.

- In V. L. Patel & G. J. Grogen (Eds.), *Proceedings of the 10th Annual Conference of the Cognitive Science Society* (pp. 139-145). Hillsdale, NJ: Lawrence Erlbaum.
- Schmidt, H. G., Norman, G. R. & Boshuizen, H. P. A. (1990). A cognitive perspective on medical expertise: Theory and implications. *Academic Medicine*, 65, 611-621.
- Schön, D. A. (1983). *The reflective practitioner: How professionals think in action*. New York: Basic Books.
- Sibert, L., Charlin, B., Corcos, J., Gagnon, R., Grise, P., & van der Vleuten, C. (2002). Stability of clinical reasoning assessment results with the script concordance test across two different linguistic, cultural and learning environments. *Medical Teacher*, 24, 537-542.