Script questionnaires: their use for assessment of diagnostic knowledge in radiology

B. CHARLIN', C. A. BRAILOVSKY[2], L. BRAZEAU-LAMONTAGNE', L. SAMSON[3], C. LEDUC' & C. VAN DER VLEUTEN'

'University of Sherbrooke, [2]Laval University, 'University of Montréal, Canada; 'University of Maastricht, The Netherlands

SUMMARY *The Diagnostic Script Questionnaire (DSQ) is innovative in its format and its scoring process. It stems from the hypothetico-deductive reasoning and the illness script theories. In this exploratory study arguments favouring a progressive construction of illness scripts along with clinical experience were sought and the possible use of the script concept was tested to assess clinical competence. The research was conducted in radiology. The investigation used the radiological hyperlucent unilateral hemithorax syndrome as diagnostic challenge. The questionnaire was answered by three groups of participants: nine seasoned radiologists (with at least three years of experience), seven radiology residents (from year I to year S), and 14 clerkship students who completed the questionnaire at the end of their radiology rotation. The data support the hypothesis of a progressive construction of specialized memory structures from students to seasoned radiologists. The DSQ was able to discriminate among participants according to their expected level of diagnostic clinical competence.*

Introduction

According to the cognitive model of the mind, knowledge is structured in memory in meaningful ways, and reasoning in a domain depends upon the structure of knowledge in memory (Irby, 1997). Hence, expertise built up can be characterized by the development of idiosyncratic memory structures, which consists in meaningful sets of connections among abstract concepts and/or specific experiences (Regehr & Norman, 1996). These memory structures are adapted to handle the tasks of the domain, so the development of expertise relates to a reorganization of knowledge in order to do these tasks (Feltovich, 1983).

According to a classic model of medical diagnostic reasoning, physicians use a hypothetico-deductive process (Elstein *et al.,* 1978; Barrows *et al.,* 1982). After collecting a few initial cues, they generate a set of competing relevant hypotheses and activate the knowledge associated with these hypotheses. In doing so, they actively look for present or absent signs that will rule in or rule out the hypotheses. This means that the capacity to come to a final diagnosis is related, on the one hand, to factual knowledge about signs and symptoms but, on the other hand, is even more related to the knowledge of relationships between these signs and

symptoms for relevant competing diagnostic hypotheses (Charlin, 1994). Hence, from a cognitive perspective, the acquisition of a networked structure of knowledge is of paramount importance in the construction of the capacity to make adequate diagnoses. Grant & Marsden (1983) have demonstrated that the more structured the memorized knowledge, the greater the clinical expertise. Then, Schmidt *et al.* (1990) proposed that the knowledge structures physicians use in clinical situations be referred to as illness scripts, and that such specialized knowledge structures begin to appear only when students are exposed to their first genuine diagnostic situations.

In the present study we looked for arguments supporting the idea that the construction of illness scripts progresses with clinical experience and we explored the possible use of the script concept to assess clinical competence. The study was carried with the assumption that a questionnaire putting clinicians in specific contexts and probing their interpreting capacities when faced with activated hypotheses will explore the relations between elements of actual clinical knowledge, i.e. will disclose evidence of elaborated knowledge that is considered by Coles (1990) and Bordage (1994) as a key for expertise.

The hypothetico-deductive reasoning process comprises a script activation phase followed by data collection and data interpretation phases. In our view, script activation takes place by a process that may belong to pattern recognition. Since script questionnaires are designed to skip the script activation phase (relevant hypotheses are specifically provided) and that of data collection (data are provided in the questions), they specifically focus on data interpretation. Hence they evaluate the component of elaborate clinical knowledge needed to confirm or reject activated hypotheses.

In radiology, a visual domain, specialists use the hypothetico-deductive reasoning process (Lesgold *et al.,*

Correspondence: Professor B. Charlin, Director, Unit of Research and Development in *Medical Education, Faculty of Medicine, University of Montreal,* E-mail: charlinb@meddir.umontreal.ca

1982; Norman *et al.,* 1992). It is interesting to explore visual domains (e.g. radiology) because data can be accessed at once, while in most other clinical contexts data are available sequentially. Hyperlucent unilateral hemithorax was selected among classical radiology syndromes for its rather wide range of differential diagnoses. Three groups of different levels of clinical experience were tested: seasoned radiologists, residents and clerkship students (at the end of a rotation in radiology). Two research hypotheses are tested in this work. (1) A questionnaire exploring the organization of clinical knowledge rather than its accumulation will provide scores reflecting differences of skills and experience among groups. Scores will also have enough variability and discriminant power to consider the questionnaire as a possible tool for assessment of clinical competence. (2) Because of less organized knowledge, less experienced clinicians will perceive each piece of radiological information as independent from the others. Consequently, the item/participants interaction should decrease with the increase of expertise while the importance of individual item difficulties in the total score increases with the increase in clinical expertise.

Methodology

Diagnosis script questionnaires (DSQs) are built according to the following steps:

1. The questionnaire depicts one (or several) clinical situation(s) within a short vignette.

2. Experts are asked which signs (positive and negative) should be looked for in such a situation, and what are the relevant hypotheses (the differential diagnosis). These hypotheses are then specified within the questionnaire.

3. Questions are written into the model (Table 1): if you are thinking of hypothesis A and you discover a sign Z, what is the effect on your hypothesis? Answers are placed on a seven-point Likert scale, with values ranging from (A) `it definitely rejects the hypothesis' to (G) `it can only be that hypothesis'. The midpoint of the scale (D) stands for `there is no effect on the hypothesis'.

The clinical situation was a left hyperlucent hemithorax on a chest X-ray (PA incidence) for which four diagnoses (the more relevant competing hypotheses) were to be con

sidered: pneumothorax, mastectomy, compensatory hyperinflation, and air trapping. A radiologist generated the positive or negative signs to be sought in this situation. For each hypothesis there were less than 10 positive signs. Negative signs were positive signs of the other competing hypotheses. Hence, a questionnaire of 49 questions was then constituted (Table 1).

The questionnaire was administered to three groups of participants: nine seasoned radiologists (with at least three years of experience), all residents (seven, from year 1 to year five) in the radiology program at Sherbrooke University, and all clerkship students (14) completing a onemonth elective rotation in radiology in the academic year 1994-95. All subjects were volunteers. All questionnaires were answered fully with the exception of one question in one questionnaire.

The nine experts were asked whether questions reflected their radiological practice, and their answers to the questions were used to construct the marking grid. The scoring process is innovative. The score for each question is the proportion of experts who gave the same answer for the question. For instance, if on a question 7/9 experts answered (F), and 2/9 answered (E), scores would be 0.77 (i.e. 7/9) for (F) answers, 0.22 (i.e. 2/9) for (E) answers; all other answers are scored at 0. Statistics were computed with the global scores obtained by adding the score obtained on each item for each participant. In this exploratory research, faculty members were considered as the criterion group and at the same time were included in the comparison among participants.

Three different types of statistical analyses were performed. (1) Descriptive statistics of the participants' scores on the DSQ, followed by a factorial analysis of variance (ANOVA) to study differences between groups' means. For all the studies, the homogeneity of group variances was estimated with Levene's test. (2) Item analysis (i.e. the study of reliability coefficients, item/total correlations, and alpha values if the item is excluded from the total) were performed using the SPSS system for Macintosh. (3) Generalizability studies were performed using the Etudgen program developed by McNicoll *et al.* (1996). The facets for the analysis were participants and items. The generalizability coefficients were calculated using the persons X items design (P X I). For the analysis of the interactions between persons and items, facet analyses were performed using the items as the differentiation facet and the participants as the instrumentation facet (UP).

Table 1. Examples of items within the questionnaire.

If you find	while you were thinking of the following hypothesis	it has the following effect (please circle your answer)				
(1) Oblong peripheral air bubble	Left pneumothorax	A B	C	D	E F	G
(2) Right mediastinal shift	Left air trapping	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,

Table 2. Descriptive statistics.

Groups	n	Mean	SD	Min.	Max.	Range
Faculty	9	50.6	3.2	44.0	54.4	10.4
Residents	7	42.7	4.9	36.3	50.3	14.0
Students	14	38.3	7.8	20.6	51.5	30.9

Results

The mean scores of the three groups of participants are given in Table 2. The students show the widest range of scores, ranging from 20.6 to 51.5 (30.9), followed by the residents 36.3 to 50.3 (14) and faculty participants 44.0 to 54.4 (10.4). One student scored very high-among the scores of seasoned clinicians. After a second look that student was revealed to have been veterinarian before entering medical school. His former training included animal radiology training. Therefore he was misclassified among students and was removed from that group for the rest of the analysis.

L,evene's test of homogeneity of variance was used to verify whether the three group variances were equal. The results were not significant $(F = 1.207, p < 0.314)$, thus indicating that the three variances can be considered equal. The factorial ANOVA tested the mean groups differences with the Tukey-Kramer *post hoc* correction. There were significant differences between student and faculty groups, resident and faculty groups and student and resident groups.

The distributions of scores for individual items for the three groups of participants varied from item to item. The Cronbach alpha coefficient for the entire group was 0.83. The item total correlations showed in the majority of cases a strong and significant correlation between all items and the results of the test for the three groups. The alpha values obtained when each item was excluded from the test indicated that almost all the items contributed positively to total reliability.

The generalizability studies have shown that the variance component for the interaction between participants and items accounted for 63.5%, 56.0% and 37.7% of the total variance in the case of students, residents and faculty respectively. The variance components for items (related to item difficulties) were 36.5%, 44.0% and 62.3% for students, residents and faculty groups respectively (Figure 1). These values indicate that item difficulties and item

specificity were different for each group of participants depending on their respective expertise. Similar results were described by Bain & Pini (1996) who studied the behaviour of different student populations in Switzerland.

Discussion

Despite the limitations of an exploratory study with its inherent small cohort of participants, several results emerged. Comparing groups shows that the greater the experience in radiology, the higher the total score. This result contrasts with `the intermediate effect' found in most research concerning assessment of competence: experienced clinicians score hardly better or even worse than end-of-training residents (van der Vleuten, 1996). An explanation may be that most assessment tools, and especially written ones, probe factual knowledge and interpretation of data using factual knowledge. Script questionnaires go further. They explore the weighing data capacity in the making of clinical decision, i.e. clearly a skill belonging to clinical competence. Therefore DSQ can discriminate between individuals according to their clinical experience. Our results also favour the idea that during medical training, organized knowledge structures adapted to clinical tasks-progressively build up. Analysis for construct validity is a never-ending process (Cronbach & Meehl, 1955), but our observations add to those provided by the series of experimentation done by Custers (1995, Custers *et al.,* 1996).

Two individual results warrant some comments. First, one student scored very high, scoring among the range of seasoned clinicians. Investigation of his training background revealed that he was more trained than the other students in the group, and for the reason we decided to remove him from that group for subsequent inference statistical analyses. Second, a seasoned clinician had a relatively low score. We found no obvious explanation, but one has to remember that due to our scoring process, the
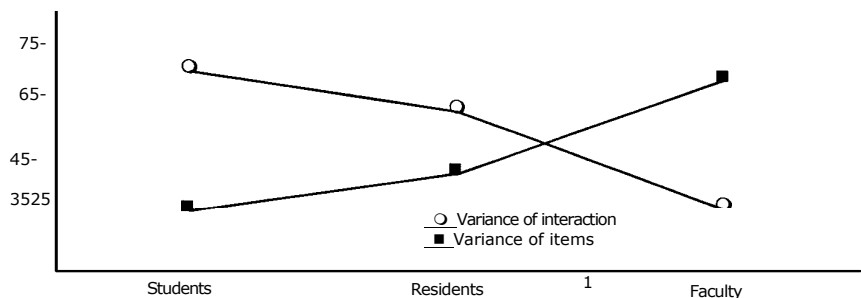
Figure 1. Percentage of variance for item/participants interaction and for item difficulty for each group.

DSQ gives high marks when answers are similar to those of the majority of the experts who are asked to build the scoring system. One interpretation of this result might be that some clinicians have a diagnostic reasoning process that differs from that of the majority. Such interpretation warrants subsequent research on that theme, knowing that one in nine experts is a significant ratio.

The reliability of the DSQ appears to be high, with a value of 0.83 for the entire group Cronbach alpha coefficient. The DSQ also shows variability and discriminant power, with score distribution for each item for the three groups of participants varying from item to item. This was due to the relative difficulty of the single item for the different subjects. Furthermore, individuals who had the same total score showed high variations in their scores for individual items, a phenomenon called case specificity, that is, the low predictability of the performance on one item as against the performance on another item.

The generalizability studies were utilized to test our second research hypothesis: the variance of the interaction between items and participants decreases when clinical experience increases, while the importance of individual item difficulty in the total score increases when clinical expertise increases. In other words, because of less organized knowledge, the less experienced clinicians will perceive each piece of clinical information as independent from each other. Results showed that the variance of interaction decreased from the student group to the faculty group, the resident group being intermediate. The variance component for the interaction between participants and items accounted for 63.5%, 56.0% and 37.7% of the variance in the case of students, residents and faculty respectively. This variance component seems to be related to the structure of the participants' knowledge. The less experienced perceive each item as a unique task whereas the experts have global perception across items, the residents being in between. On the other hand, the variance components for items were 36.5%, 44.0% and 62.3% for students, residents and faculty groups respectively. These values indicate that item difficulties and item specificity were different for each group of participants depending on their respective expertise. All these results favour our second research hypothesis.

The exploratory format of the study carries its own limitations, mainly due to the small numbers of participants. Nevertheless they correlate well with another study done in obstetrics, with a larger number of participants (Charlin *et al.,* 1998). Both studies support the concept, introduced by Grant & Marsden (1983), of an acquisition of expertise based on the progressive construction of specialized knowledge structures. Our results also raise arguments for the use of DSQ as an assessment tool of clinical competence:

(1) Scores have a large variability, so it should be possible to discriminate among individuals according to their clinical competence.
(2) Participants made the comment that DSQ was interesting to complete because questions are similar to those they ask in real settings. This indicates that DSQ might have a positive steering effect on learning.
(3) Clinical competence is a multitrait construct (van

derVleuten, 1996) and this is true even for diagnostic competence.

The hypothetico-deductive reasoning process is made up of at least three different phases, and each of them could be tested in a comprehensive assessment of the diagnostic process: activation of a set of relevant competing hypotheses, collection of data oriented by the hypotheses and interpretation of the data to confirm or reject these hypotheses. In this study we focused on the interpretation phase only. Other questionnaires based on similar principles could be constructed to assess the other phases. Also other tasks belonging to clinical competence, such as the management of patients, should be addressed [1].

Note
[1] We are already pursuing research on the use of the DSQ in radiology with two developments in the questionnaire: (1) introduction of actual X-ray readings, to increase test validity, and (2) inclusion of several vignettes instead of one to enlarge sampling of situations in the test.

Notes on contributors

BERNARD CHARLIN is Professor of Surgery. He is now Director of the unit of Research and Development in Medical Education University of Montreal.

CAaLOS BRAILOVSKY is Director of the 'Centre d'évaluation dans les sciences de la santé', Faculty of Medicine, Laval University.

Lucir BaAZEAU-LnIVtoNTAGNE is Professor of Radiology and Associate Dean, Faculty of Medicine, University of Sherbrooke.

LouisE SAMSON is Professor of Radiology, Faculty of Medicine, University of Montréal.

CHARLES Lenuc is Associate Professor of Family Medicine with cross-appointment in Radiology, Faculty of Medicine, University of Sherbrooke.

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

References

BAIN, D. & PtNt, G. (1996) *Pour évaluer vos évaluations: La généralisabilité, mode d'emploi* (Genoa, Centre de recherches psychopédagogiques).

BARROw5, H.S., NORMAN, G.R., NEUFELD, V.R. & FEIGHTNER, J.W. (1982) The clinical reasoning of randomly selected physicians in general medical 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.

CHARLIN, B. (1994) Le schéma comme structure de connaissances sous-jacente aux hypothèses dans l'investigation clinique médicale, mémoire pour l'obtention du diplôme de maîtrise ès arts en sciences de l'éducation, Université de Sherbrooke.

CHARLIN, B., BRAILOVSKY, C., LEDUC, C. & BLOUIN, D. (1998) The diagnosis script questionnaire: a new tool to assess a specific dimension of clinical competence, *Advances in Health Science Education,* accepted for publication.

COLIS, C.R. (1990) Elaborated learning in undergraduate medical education, *Medical Education,* 24, pp. 14-22.

CRONBACH, L.J. & MsI:HL, P.C. (1955) Construct validity in psy chological tests, *Psychological Bulletin,* 52, pp. 281-302. CusTr:RS, J.F.M. (1995) The development and function of illness scripts. Studies on the structure of medical diagnostic knowledge, PhD thesis. Maastricht, Netherlands, Universitaire Pers Maastricht.

Cusrt.RS, J.F.M., REGEHR, G. & NORMAN, G.R. (1996) Mental representations of medical diagnostic knowledge: a review, *Academic Medicine,* 71, pp. S55-S61.

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

Fr:LTOVICH, P.J. (1983) Expertise: reorganizing and refining knowl edge for use, *Professions Education Research Notes,* 4, pp. 5-9.

GRANT, J. & MARSnt:N, P. (1988) Primary knowledge, medical edu cation and consultant expertise, *Medical Education,* 22, pp. 173 179.

IRBY, D.M. (1997) Editorial, *Academic Medicine, 72, p.* 116.

LESGOLD, A., RvBtNSON, H., FELTOVICH P. *et al*. (1988) Expertise in a complex skill: diagnosing X-ray pictures, in: M. T. H CHI, R.

Gt.asl:x & M. J. FARE (Eds) *The Nature of Expertise, pp.* 311-342 (Hillsdale, NJ, Lawrence Erlbaum).

McNICOLL, A., BRAIIAVSKY, C.A., BERTRAND, R. & CARDINET, J. (1996) EtudGen, programme pour l'analyse de la généralisabilité pour Macintosh, © CESSUL 1992, 1996, in: D. BAIN & G. PINI *Pour évaluer vos évaluations: La généralisabilité, mode d'emploi, p. 51* (Geneva, Centre de recherches psychopédagogiques).

NORMAN, G., COBLENTZ, C., BROOKS, L. & BADCOCK, C. (1992) Expertise in visual diagnosis: a review of the literature, *Academic Medicine,* 67, pp. S78-S83.

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

SCHMIDT, H.G., NoRMAN, G.R. & BosttuIZBN, H.P.A. (1990) A cognitive perspective on medical expertise: theory and implications, *Academic Medicine,* 65, pp. 61 1-621.

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.