

## The Diagnosis Script Questionnaire: A New Tool to Assess a Specific Dimension of Clinical Competence

B. CHARLIN<sup>1</sup>, C. BRAILOVSKY<sup>2</sup>, C. LEDUC<sup>1</sup> and D. BLOUIN<sup>1</sup>

<sup>1</sup>Faculté de médecine, Université de Sherbrooke, Québec, Canada; <sup>2</sup>CESSUL, Faculté de médecine, Université Laval, Québec, Canada

**Abstract.** The Diagnosis Script Questionnaire (DSQ) assesses a specific skill of clinical competence: the ability to weigh collected information in light of entertained hypotheses. The questionnaire presents a clinical vignette for which several hypotheses are relevant. The model of the questions is: if you are thinking of hypothesis A and you find sign Z, what is the effect on your hypothesis? Answers are placed on a 7-point Likert scale, with values ranging from “it can only be this hypothesis” to “this hypothesis is definitely rejected.” The scoring process is innovative and reflects the variability of answers among experts.

The questionnaire was administered in gynecology-obstetrics; 103 respondents, divided into three groups, 15 faculty, 12 residents, and 76 clerkship students volunteered. Mean global scores were 45.3 for faculty, 40.5 for residents, and 35.8 for students. The differences between the three groups were statistically significant with Welch ANOVA ( $p < 0.001$ ). The Bonferroni post-hoc correction however indicated that the only significant difference was between student and faculty groups ( $p < 0.001$ ). Cronbach's alpha was 0.822 for the total group; for the student and resident groups, 0.794 and 0.812 respectively. The proportion of the total variance explained by the interaction items/participants as estimated by generalizability was 42.1%, 65.4% and 73.4% for the faculty, resident and student groups respectively.

Results agree with the theories of development of clinical competence which states that knowledge structures specifically adapted to diagnostic tasks appear with clinical experience. This new assessment tool appears promising and warrants future development.

**Key words:** clinical reasoning, evaluation, knowledge structures, medical students, physicians, residents

### Introduction

A new assessment tool, the Diagnosis Script Questionnaire, (DSQ) has been designed to assess a specific skill of clinical competence: the ability to weigh clinical information in light of entertained hypotheses in diagnostic situations. The DSQ is grounded in two theories of clinical reasoning: the hypothetico-deductive theory and the illness script theory.

Since the work by Elstein and collaborators in 1978, we know that clinical reasoning is a hypothetico-deductive process characterized by the early generation of hypotheses, oriented data collection and decision making judgment, and the use of data to confirm or reject hypotheses. According to Schmidt, Norman and Boshuizen (1990), the acquisition of clinical expertise is related to the development of specialized knowledge structures, referred to as illness scripts, that contain

<i>If you find</i>	<i>while you were thinking of the following hypothesis</i>	<i>it has the following effect (please encircle your answer)</i>
1. Bleeding happening at 39 weeks of pregnancy	Placenta previa	A B C D E F G
2. Candida vaginitis in early pregnancy	Abruptio placenta	A B C D E F G
3. Placenta normally inserted at 20 weeks of pregnancy	Uterine rupture	A B C D E F G
4. Chronic hypertension diagnosed before pregnancy	Abruptio placenta	A B C D E F G
5. Cervical conization done one month before pregnancy	Uterine rupture	A B C D E F G
6. Late deceleration of the fetal heart rate	Placenta previa	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*

Figure 1. Principle of the questionnaire.

the clinically relevant information that clinicians will use in their clinical activities. The theory considers that in diagnostic situations clinicians bring in their working memory knowledge related to each relevant hypotheses. Such knowledge encompasses all clinical features (that is symptoms, signs, laboratory or radiology data and/or previous experiences) that clinicians know to be useful for diagnosis. Knowledge activated from memory is then used in a deductive process to actively seek information that will allow confirmation or rejection of entertained hypotheses (Charlin, 1994).

This theory rests on the assumption that clinicians have idiosyncratic memory structures which are meaningful sets of connections among clinical features (Regehr and Norman, 1996) and that these connections reflect the organization of knowledge that takes place when students begin to be exposed to real clinical tasks (Schmidt, Norman and Boshuizen, 1990). The concept of illness script has been extensively tested in a recent series of experiments by Custers (1995).

The DSQ (Figure 1) is built around the information clinicians look for in the hypothetico-deductive processes of clinical reasoning. The questionnaire depicts within a short vignette a clinical situation for which a few hypotheses are relevant, all of them being specified. A series of questions are presented, based on the model: 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 7-point Likert scale, with values ranging from "it can only be that hypothesis" to: "it definitely rejects the hypothesis". The midpoint of the scale stands for "there is no effect on the hypothesis."

Research hypotheses were: (1) Since we postulate that specific knowledge structures are built with clinical experience, if the group of higher expertise is considered as a reference group, scores reflecting the ability to weigh clinical information will be lower for the other groups; (2) With this assessment tool it

should be possible to discriminate students with different levels of competence; (3) The interaction item/candidates increases with the decrease of expertise while the importance of individual item difficulties in the total score increases with the increase of the clinical expertise.

### **Material and Method**

This new kind of questionnaire was tested in the field of obstetrics. The clinical situation was bleeding in the third trimester of pregnancy. The four more relevant competing hypotheses (placenta previa, abruptio placenta, uterine rupture, and bleeding from cervical os) were explicitly stated. An obstetrician generated the positive and negative signs he would seek in this situation. For each hypothesis there were less than 10 positive signs; negative signs were the positive signs of the other competing hypotheses. A questionnaire of 50 questions was then drawn up. The resulting questionnaire reflects the clinical reasoning of one expert. The questionnaire was then shown to two other experts for validation before use. This validation resulted in only minor wording corrections.

The study ( $n = 103$ ) compared three groups with different levels of clinical experience: faculty members ( $n = 15$ ), from the gynecology-obstetrics departments of two universities (University of Sherbrooke, and Laval); all the residents from the Gynecology/Obstetrics residency program (from year 1 to year 5) ( $n = 12$ ) that were in training at Sherbrooke Medical School, and all clerkship students within a class in Sherbrooke just before the end of their MD program ( $n = 76$ ). Students and residents from Sherbrooke do not differ in training from those of the other medical school in Canada, and they perform in a similar way in national examinations. Ten students refused to answer the questionnaire. Due to anonymity of participation, we do not have any data allowing to compare them to those who have accepted.

The scoring process is an original aspect of this questionnaire. We approached it with the idea that any faculty member's answer reflects the opinion of an expert, and that we should not discard answers for which there is no agreement among all the experts. In other words, we thought that any answer given by an expert has an intrinsic value even if other experts do not agree with it. Hence, scores for each item were computed from the frequencies given to each point of the Likert scale by faculty members. If for instance, for a given question, the distribution of faculty members answers is 11 for one point of the scale and 4 for another one, scores for the participants who answer that item in the same manner would be 0.73 (11/15), 0.27 (4/15) and 0 for all others. The results of the test are represented by the sum of the scores obtained at each question.

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 using the Levenes' test. Pairwise comparisons were then used to determine precisely which differences of

Table I. Comparison of mean scores by groups

Groups	N	Mean	St. dev.	Min	Max
Faculty	15	45.3	3.7	36.3	52.4
Residents	12	40.5	7.1	28.5	49.5
Students	76	35.8	7.4	13.3	48.7

Students vs. faculty:  $p < 0.001$  (Welch Anova and Bonferroni post-hoc correction); other comparisons between groups were not significant.

scores between groups of participants were significant. Since the study required multiple contrast analyses, the Bonferroni correction was utilized to decrease the probability of finding significant results due to chance. (2) Item analysis (i.e., the study of reliability coefficients of the test for each group, item/total correlations, and alpha values if the item is excluded from the total) were performed using the SPSS system for Macintosh. (3) Generalisability studies were performed using the Etudgen program developed by McNicoll et al. The facets for the analysis were participants and items. The generalisability 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 differentiation facet and the participants as instrumentation facet (I/P).

Statistics were computed with the global scores obtained by adding for each participant the score obtained on each item. In this exploratory research, faculty members are considered as a reference group (but not as a gold standard).

## Results

The mean scores of the three groups of participants are shown in Table I. The students show the widest range of scores, ranging from 13.3 to 48.7 (36), followed by residents (21) and faculty participants (16). One faculty participant behaved in a different manner than his colleagues, his score being below the 10th percentile of his group. From these observations and the scores' standard deviations, it appeared that the groups were not homogeneous. To test this hypothesis the Levene's test of homogeneity of variance was used to verify whether the three group variances were equal. The results were significant ( $F = 4.2101$ ,  $p < 0.018$ ), thus indicating that the three variances were in fact not equal. The Welch ANOVA test for unequal variances was then used to compare groups' means. The results were significant for the three groups ( $p < 0.001$ ). The Bonferroni post hoc correction however, indicated that the only significant difference was present between the scores of student and faculty groups ( $p < 0.001$ ). Furthermore, even if the mean score of residents appears to be different from these of the faculty participants, the lack of significance in the difference of the scores between residents and faculty can be explained by the higher variability of residents scores.

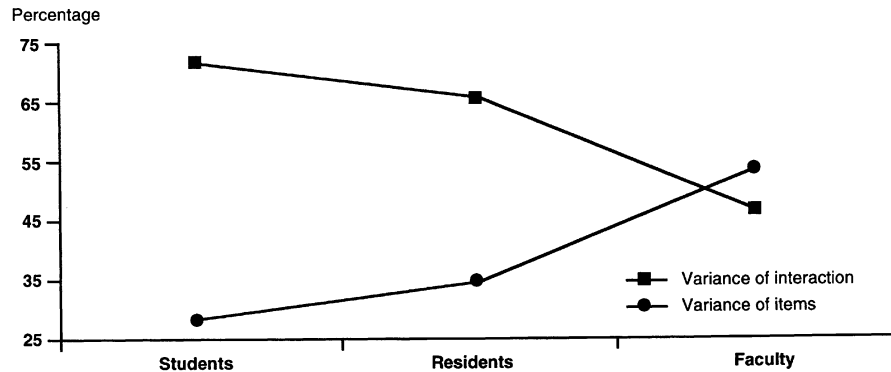


Figure 2. Percentage of variance for item/candidates interaction and for item difficulty for each group.

The distributions of scores for individual items for the three groups of participants varied from item to item. This was due to the relative difficulty of the individual items 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 of one item from the performance on another item.

The alpha coefficient for the entire group was 0.822. The alpha coefficient for the residents' group was 0.812 and for the students group 0.794. The elimination of the "low score" participant in the faculty group increased the mean and decreased the standard deviation of the group but had no influence on group test reliability and it was decided to include the scores of this participant in the subsequent analyses. 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 resident and student 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 true variance increased from the faculty group to the student group, the resident group being intermediate (Figure 2). The variance component for the interaction between participants and items (Bain, D. and Pini, G., 1996) accounted for 71.6%, 65.4% and 46.3% 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 ones perceive each item as an unique task whereas the experts have global perception across items, the residents being in between. The variance components for items was 53.7% for faculty participants, 34.6% for residents and 28.4% for students. These values indicate that item difficulties and item specificity were different for each group of participants depending on their respective expertise.

## Discussion

Contemporary methods of assessment of clinical competence have repeatedly shown the puzzling fact that experienced clinicians score little better and sometimes worse than less experienced clinicians or students (Van der Vleuten, 1996; Newble 1982; Marchall, 1977). Most of these methods of assessment measure clinical factual knowledge rather than the organization of knowledge that allows clinicians to recognize and handle situations effectively. According to Feltovich (1983), the development of expertise is largely a matter of reorganizing knowledge and cognitive processes to fit the demand of tasks within the domain of expertise. An assessment tool that would explore the ability among clinicians to weigh incoming clinical information to make judgments about entertained hypotheses may support what Bordage (1994) and Coles (1990) call elaborated knowledge which is supposed to be a key component of clinical competence. It is in this perspective that we developed the DSQ.

In an usual clinical situation, a skilled clinician activates relevant diagnostic hypothesis by pattern recognition. He then uses illness scripts to ask questions and perform physical examinations in order to make decisions about the right diagnostic. In the DSQ the pattern recognition stage is skipped because the relevant hypotheses are given. The goal of the questionnaire is to assess if the richness of existing illness scripts allows adequate interpretation of provided clinical data.

The content validity of the QSD and the relevance of competing hypotheses were assessed by two gynecology-obstetric experts. The reliability (homogeneity) of the test was calculated and show acceptable coefficients. The test-retest reliability was not yet assessed because the QSD was utilized only once with each group. However, the instrument is being tested in other settings in various medical specialties and test-retest is in the way.

Our results show an increase in the mean scores on the DSQ of groups with different clinical expertise, the less experienced getting the lower results. This may mean that the DSQ measures a dimension for which, as one should expect, experts get better scores than less experienced subjects, thus supporting the construct validity of the instrument. Furthermore, the observation that the interaction item/candidates increases inversely to the expertise of the group under study, whereas the importance of the item difficulty decreases, suggests that the expertise of the individuals plays a role in overcoming the problems raised by the content of the items.

Clinical competence is a multidimensional entity. No single assessment method can adequately measure it. The DSQ measures a specific dimension of clinical competence. It is built by interviewing clinicians about the information they look for in given clinical situation and is relatively easy to build and administer. The large variability found among students suggests that it may be possible to discriminate among students with the use of the DSQ as an assessment tool. To validate this possibility, we are now analyzing and comparing the results of strong and weak

students' scores on the DSQ to their performances on other already validated tests. The results obtained with participants in obstetrics have to be confirmed with other studies that will allow the verification of the validity of this model of evaluation. In our experimental design we used only one clinical situation within one medical discipline. We are already doing similar research in other disciplines, with more clinical situations.

In medical education few assessment tools are available that allow the comparison of clinical knowledge among subjects as different as students, residents and faculty. Depending on the goal, in DSQ construction it is possible to introduce items that could differentiate between the groups of candidates in relation to their level of clinical competence. If one wants to build a questionnaire that assesses clinical competence related to the more frequent clinical situations of a domain, a ceiling effect will be reached between senior residents and faculty. On the contrary, it is possible to build a questionnaire that would allow a good dispersion of candidates' results but this will need the introduction of items that reflect competence in a rather rare clinical situation and hence this will pose a problem of content validity of the instrument.

Results agree with theories that relate the development of clinical competence with the acquisition of specific knowledge structures adapted to clinical tasks (Schmidt and Norman 1990; Feltovich 1983; Bordage 1994). However, the fact that one expert gave very different answers from other experts may indicate that a few individuals have a different way of reasoning in diagnosis situations and open further research in the definition of "expert competence" for this kind of research. Finally, our results showed that experts provide different answers for questions that do not appear ambiguous. This raised questions about the usual practice of seeking consensus among experts in the scoring process of examinations. Consensus decisions (i.e. to decide in an examination which are the accepted answers) are classically obtained among experts in group discussions. From the results presented, when they make individual choices, the variability of decisions is quite high. It may be more valid to permit students the same differences among answers that experts themselves provide for the very same questions.

### **Acknowledgments**

This project has benefited of a grant from the Association of Canadian Medical Schools and the Medical Research Council of Canada.

### **References**

- Bain, D. & Pini, G. (1996). *Pour évaluer vos évaluations: La généralisabilité, mode d'emploi*. Centre de recherches psychopédagogiques: Genève.
- Bordage, G. (1994). Elaborated Knowledge: A Key to Successful Diagnostic Thinking. *Academic Medicine* **69**: 883-85.

- 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.
- Coles, C.R. (1990). Elaborated Learning in Undergraduate Medical Education. *Medical Education* **24**: 14–22.
- Custers, J.F.M. (1995). *The Development and Function of Illness Scripts. Studies on the Structure of Medical Diagnostic Knowledge*. PhD thesis. Universitaire Pers Maastricht, Maastricht, Netherlands.
- Elstein, A.S., Shulman, L.S. & Sprafka, S.A. (1978). *Medical Problem Solving: An Analysis of Clinical Reasoning*. Harvard University Press: Cambridge, MA.
- Feltovich, P.J. (1983). Expertise: Reorganizing and Refining Knowledge for Use. *Professions Education Research Notes* **4**: 5–7.
- Marshall, J. (1977). Assessment of Problem-Solving Ability. *Medical Education* **11**: 329–334.
- McNicoll, A., Brailovsky, C.A., Bertrand, R. & Cardinet, J. (1996). EtudGen, programme pour l'analyse de la généralisabilité pour Macintosh. © CESSUL 1992, 1996. In D. Bain et G. Pini "Pour évaluer vos évaluations: La généralisabilité, mode d'emploi". Centre de recherches psychopédagogiques, Genève. p. 51.
- Newble, D.I., Hoare, J. & Baxter, A. (1982). Patient Management Problems: Issues of Validity. *Medical Education* **16**: 137–142.
- Regehr, D. & Norman, G.R. (1996). Issues in Cognitive Psychology: Implications for Professional Education. *Academic Medicine* **71**: 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**: 611–21.
- Van der Vleuten, C.P.M. (1996). The Assessment of Professional Competence: Development, Research and Practical Implications. *Advances in Health Sciences Education* **1**: 41–67.