Exploring how students think: a new method combining think-aloud and concept mapping

Pierre Pottier,[1] Jean-Benoit Hardouin,[2] Brian D Hodges,[3,4] Marc-Antoine Pistorius,[1] Jérome Connault,[1] Cécile Durant,[1] Renaud Clairand,[1] Véronique Sebille,[2] Jacques-Henri Barrier[1] & Bernard Planchon[1]

OBJECTIVES A key element of medical competence is problem solving. Previous work has shown that doctors use inductive reasoning to progress from facts to hypotheses and deductive reasoning to move from hypotheses to the gathering of confirmatory information. No individual assessment method has been designed to quantify the use of inductive and deductive procedures within clinical reasoning. The aim of this study was to explore the feasibility and reliability of a new method which allows for the rapid identification of the style (inductive or deductive) of clinical reasoning in medical students and experts.

METHODS The study included four groups of four participants. These comprised groups of medical students in Years 3, 4 and 5 and a group of specialists in internal medicine, all at a medical school with a 6-year curriculum in France. Participants were asked to solve four clinical problems by thinking aloud. The thinking expressed aloud was immediately transcribed into concept maps by one or two 'writers' trained to distinguish inductive and deductive links. Reliability was assessed by estimating the inter-writer correlation. The calculated rate of inductive reasoning, the richness score and the rate of exhaustiveness of reasoning were compared according to the level of expertise of the individual and the type of clinical problem.

RESULTS The total number of maps drawn amounted to 32 for students in Year 4, 32 for students in Year 5, 16 for students in Year 3 and 16 for experts. A positive correlation was found between writers (R = 0.66–0.93). Richness scores and rates of exhaustiveness of reasoning did not differ according to expertise level. The rate of inductive reasoning varied as expected according to the nature of the clinical problem and was lower in experts (41% versus 67%).

CONCLUSIONS This new method showed good reliability and may be a promising tool for the assessment of medical problem-solving skills, giving teachers a means of diagnosing how their students think when they are confronted with clinical problems.

*Medical Education 2010: **44**: 926–935*
doi:10.1111/j.1365-2923.2010.03748.x

[1]Department of Internal Medicine, Nantes University Hospital Centre, Nantes, France
[2]Team for Biostatistics (EA 4572), Department of Clinical Research and Subjective Measures in Health Science, University of Nantes, Nantes, France
[3]Wilson Centre for Research in Education, University of Toronto, Toronto, Ontario, Canada
[4]Department of Psychiatry, University of Toronto, Toronto, Ontario, Canada

Correspondence: Dr Pierre Pottier, Service de Médecine Interne A, Hôtel Dieu, CHU Nantes, 44093 Nantes Cedex 1, France.
Tel: 00 33 24 00 83 352; Fax: 00 33 24 00 83 379;
E-mail: pierre.pottier@univ-nantes.fr, pier@blackberry.orange.fr

INTRODUCTION

Tutors involved in teaching clinical reasoning have few practical means of diagnosing how their students think when they are confronted with clinical problems. For a medical teacher who wants to improve problem-solving strategies, it is essential to be able to represent how students think in order to provide an appropriate diagnosis of their reasoning difficulties, to give sensible feedback and to adjust teaching goals. This research work was aimed at developing a workable method for exploring medical problem solving that would allow for the rapid identification of an individual's style of reasoning.

According to the information processing theory described by Newell and Simon,[1] several types of strategy for problem solving have been identified. Pattern recognition[2–4] is a non-analytical unconscious strategy in which the solution appears immediately because a prototype or an instance of the situation has already been memorised. Scheme-inductive process and hypothetico-deductive reasoning[4–6] are analytical conscious procedures. Inductive reasoning progresses from facts to hypotheses, whereas deductive reasoning is driven from hypotheses to facts. These strategies are activated from an initial representation of the situation built either on semantic axes, as described by Nendaz and Bordage,[7] or from a clinical script, as reported by Charlin *et al.*[8] The use of a particular strategy is highly variable from one subject to another and depends on the difficulty of the task, the question format[4] and the level of expertise of the individual.[6]

Qualitative and quantitative tools for the assessment of strategies for problem solving have been designed. Among the qualitative methods, think-aloud protocols have been widely used.[4,9–13] Verbatim transcriptions obtained with think-aloud methods have sometimes been analysed by means of quantitative scoring.[12]

Concept mapping is a quantitative tool with two main distinct scoring methods. The structural method focuses on the aspect of the map and the type of links made.[14–16] Using this method, some researchers have aimed to assess the degree of similarity between different structures of maps.[17,18] The relational method gives more importance to the value of the links (false/true) in comparison with a reference map,[15,16,19] but, as Acton and colleagues[20] have previously shown, using a reference to score a student's map may be problematic because it may lead to heterogeneous conclusions about the student's knowledge. Concept mapping has been used as a learning method,[21–25] as well as an assessment tool.[15,25,26] Concept mapping is typically used to evaluate (or to teach) factual[16,19,23,24,27] or procedural knowledge,[22] as opposed to clinical reasoning skills.[26]

Research on clinical reasoning based on concept mapping is faced with some difficulties. The main problem concerns the absence of a reference standard in terms of problem-solving strategies for particular clinical problems, although some strategies seem to show better accuracy in terms of diagnostic success.[4,10] Indeed, a great variety of concept maps can be expected for the same task, depending on the theoretical and practical knowledge of the individual to whom the map pertains. Moreover, regardless of the structure of the map, it is difficult to see how a piece of reasoning (and its strategies) can be decreed bad if the problem is actually solved. Secondly, writing a map while one is actively thinking about the problem to be solved might be confusing. Lastly, any large-scale use of this method is limited by the time required to learn mapping techniques. Furthermore, no scoring method has yet been designed to quantify the respective proportions of inductive and deductive procedures within clinical reasoning, although we know that solving a problem usually requires both forms of reasoning.[10,28–30] Taking these issues into account, our objective in this study was to explore the feasibility and reliability of a new method which allows for the rapid estimation of the respective contributions of inductive and deductive reasoning in medical students and experts faced with clinical problems. Such a method could prove useful in facilitating immediate diagnoses of individual students' difficulties with clinical reasoning and, consequently, could provide a basis for the provision of individual feedback. If a teacher is aware of the style of reasoning used by a particular student, he or she can then adapt the educational approach in order to progressively guide the student towards expert reasoning.

METHODS

Study design

Using a quasi-experimental quantitative study, we assessed: (i) a new method combining a think-aloud protocol and the creation of live-written maps by observers ('writers'), and (ii) a new structural method for scoring.

Study participants

Students attending a training course in the Department of Internal Medicine at a French medical school running a 6-year curriculum were informed of the study protocol and invited to enrol in the study. The anonymity of participants was guaranteed and informed consent of participants was obtained.

Three groups of medical students (selected from separate cohorts of approximately 20 students) were formed, with four students in each group. Group 1 included Year 3 students, Group 2 consisted of Year 4 students and Group 3 comprised Year 5 students. A fourth group was created to include specialists in internal medicine working in the same department, with a mean length of working experience of 4 years (Group 4).

Writing the simulated clinical problems

Four clinical problems were written based on content areas already studied during the curriculum. These problems, derived from authentic clinical cases, were designed by two medical teachers specialising in internal medicine. The clinical cases were constructed with the expectation that different types of problem-solving strategies could be used but without considering any one particular strategy as a reference standard for a given clinical problem. Appendix S1 (online) details the four problems.

- Problem 1 presented the symptoms typical of a child with an acute epiglottitis. The problem presented all the signs necessary to allow an instantaneous diagnosis using a pattern recognition strategy as described by Norman and colleagues.[2,3]
- Problem 2 presented the medical history of a 65-year-old patient with a pernicious anaemia. All the necessary signs for diagnosis were present but, because of the multi-systemic character of this disease, the correct diagnosis becomes apparent only after signs pertaining to similar syndromes have been identified. This type of analysis represents a particular form of inductive reasoning because it is driven from facts to hypotheses.
- Problem 3 presented the history of a patient complaining of jaundice. In this case, several diagnoses were plausible after analysis, whereas some other diagnoses could be excluded according to discriminating symptoms (e.g. the presence of an increased gallbladder indicated a compression of the main biliary duct and excluded hepatitis). This type of forward-driven reasoning often rests on a previously established decision-making tree and is also described as scheme-inductive reasoning.[31]
- Problem 4 presented the case of an elderly woman with a rich medical history complaining of an isolated pruritus. According to Heemskerk and colleagues,[4] we postulated that the multiplicity of possible aetiologies in this case might facilitate a systematic and a priori generation of hypotheses and consequently induce deductive reasoning (progressing from hypotheses to facts).

Study process

Each participant was invited into a room and presented by a researcher with the four clinical problems (summarised above and detailed in Appendix S1) on an A4 sheet.

The participant was asked to solve the problems one by one, by thinking aloud, and was reminded that data not reported in the wording had to be considered as absent. No time limit was imposed. The thinking expressed aloud was immediately transcribed into concept maps by one or two 'writers' who sat in the room and recorded observations directly onto maps. Two writers were used in a few groups in order to calculate inter-writer reliability. They were unable to review each other's maps before submission. If it was difficult to determine the nature of a link, the map writers were free to ask the participant to clarify his or her thinking during the session.

Writing the concept maps

Instructions for writing concept maps are given in Appendix S2 (online). Briefly:

1 concepts were linked by lines;
2 among concepts, facts were distinguished from hypotheses: all concepts reported in the wording of the clinical case were considered as facts, whereas all new concepts generated by the participant were considered as hypotheses;
3 inductive links from facts to hypotheses were drawn in red and deductive links from hypotheses to facts were drawn in green;
4 facts and hypotheses were numbered in order of appearance during reasoning, and

5 to offset the possibility that, in this context of written cases, data not present in the wording of the case might be immediately considered by participants as non-pathological (limiting the deductive checking process), reasoning was considered as deductive as soon as several hypotheses were generated at the same time, even if they were not subsequently confirmed by facts.

Writer training

In order to establish the reliability of the scores derived from the mapping method, two writers were selected from among the researchers on our team (Group W1) and from among the teachers in our faculty (Group W2). Writers in Group W1 were familiar with the theoretical framework of this study,

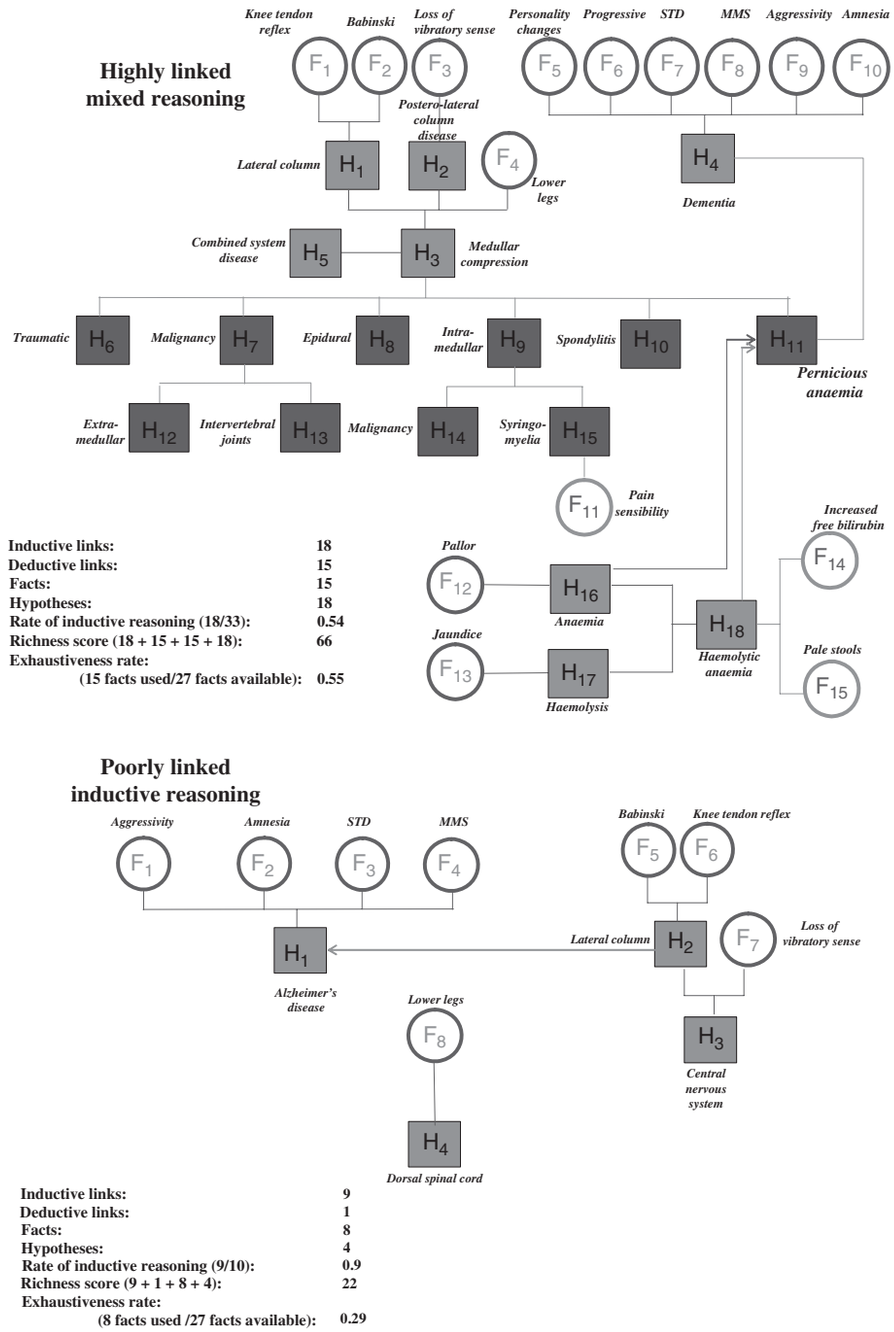

Figure 1 Example maps of highly and poorly linked reasoning and their scoring. MMS = mini-mental status; STD = spatio-temporal disorientation; F_1= fact used in position 1; H_1 = hypothesis generated in position 1

whereas writers in Group W2 were blinded to our hypotheses about the expected forms of reasoning and had no specific theoretical background in clinical reasoning other than that generated by their daily involvement in practice and teaching. Writers were trained according to the following steps:

1 the concepts of links, facts, hypotheses, inductive and deductive links were explained according to the definitions espoused above;
2 when the simulated clinical problems were presented, the writers were blinded to the type of reasoning expected for each problem;
3 scoring was demonstrated for two maps (Fig. 1), showing richly and poorly linked reasoning;
4 scores for these maps were compared with the reference score, and
5 any questions put by the writers were answered.

Writer training lasted approximately 30 minutes.

Scoring method

Numbers of inductive links, deductive links, facts and hypotheses were calculated from the maps. We also calculated a score for the richness of reasoning by creating a sum of these four parameters and computed the rate of exhaustiveness of reasoning as the ratio between the number of facts used and the total number of facts available in the case (facts considered by the authors of the cases as potentially useful for correct diagnosis). The rate of inductive reasoning was estimated from the ratio between the number of inductive links and the total number of links. The different types of reasoning expected for the four clinical problems are mapped and scored in Fig. 2. The scoring method is explained in Appendix S2 and a detailed example is given in Results.

Analysis process

The analysis was conducted in three steps.

The first step aimed to check the reliability of the scores derived from the mapping method. For Group 3 (Year 5 medical students), two maps per student were written and scored by writers in Group W1 (A and B) according to the method described above. This was repeated for Group 2 (Year 4 students) and the writers in Group W2 (C and D). Reliability was assessed by estimating: (i) the mean of the absolute difference (MAD) between the scores delivered by the two writers, and (ii) the inter-writer correlation for each scoring parameter. A large MAD value was taken to signify disagreement between the two writers, in which case their ratings were further analysed using a

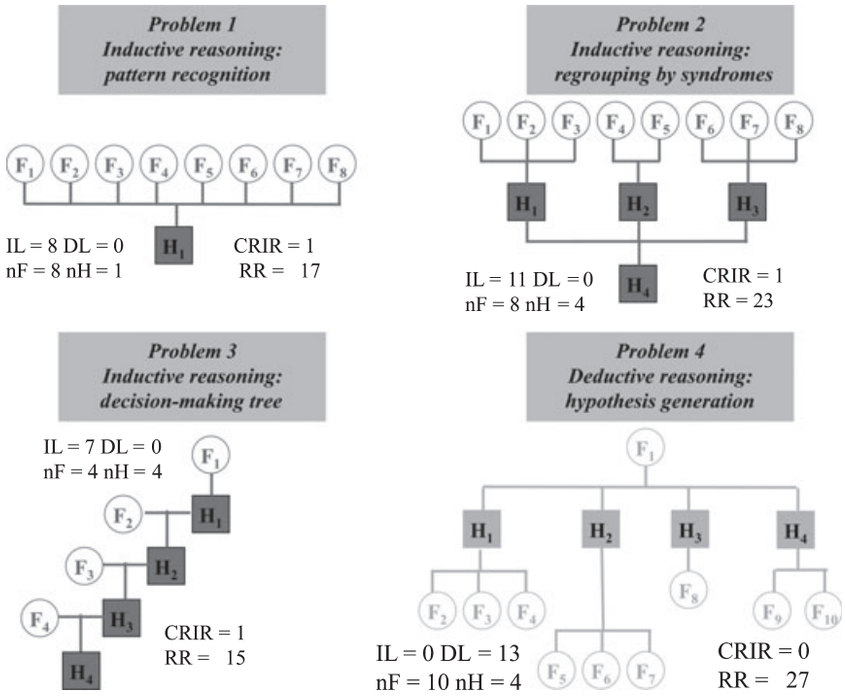

Figure 2 Reasoning expected to be induced by the four clinical problems. IL = inductive links; DL = deductive links; nF = number of facts; nH = number of hypotheses; CRIR = calculated rate of inductive reasoning (IL/[IL + DL]); RR = richness of reasoning score (IL + DL + nF + nH); F_1= fact used in position 1; H_1 = hypothesis generated in position 1

correlation. A large MAD associated with a high correlation coefficient would signify a systematic over- or underestimation of the parameter under study by one of the two writers. A large MAD associated with a weak correlation coefficient would signify a random disagreement between the two writers.

The second step aimed to estimate the validity of the method by considering variations in the rate of inductive reasoning according to the type of problem. Higher rates of inductive reasoning were expected for Problems 1, 2 and 3 than for Problem 4.

The third step aimed to compare novices and experts in terms of the richness and exhaustiveness of their reasoning (RR, ER) and in terms of the respective proportions of inductive and deductive reasoning estimated by the calculated rate of inductive reasoning (CRIR), in the knowledge that the direction of those variations would be difficult to predict given that the increase in links, nodes and concepts that has been reported to occur with increased expertise[24] has not always been confirmed.[16,19,27] This step was performed in three groups for which mapping was carried out by the same writer (thus excluding Group 2 [Year 4 students]).

Statistical analysis

Statistical analysis was performed using STATA Version 10 (StataCorp LP, College Station, TX, USA). The normality of the distribution of quantitative variables was checked by the Shapiro–Wilks test. Means were compared using paired Student's tests in cases of normality of distribution or Wilcoxon sign rank tests in cases of non-normality. Multiple means comparisons were performed by parametric or non-parametric (Kruskall–Wallis test) analysis of variance (ANOVA). In cases of significant ANOVA, multiple post hoc comparisons of means were performed using Bonferroni corrections.

The correlation between quantitative variables was estimated using Pearson's coefficient in cases of normality of distribution or Spearman's coefficient in cases of non-normality. The significance threshold was fixed at 5%.

RESULTS

General data

A total of 96 maps were drawn, including 16 by one writer for each of Groups 1 and 4, and 32 by two writers for each of Groups 2 and 3. Four participants were included in each group. On average, it took 12 minutes to complete one map. Figure 1 shows representations of highly and poorly linked reasoning with the results of scoring for each. The first map of reasoning (shown at the top of Fig. 1) was deductive and inductive (red and green links in quasi-equal proportions) and was relatively rich (richness score: 66) and exhaustive (exhaustiveness rate: 0.55) because more than half of the pertinent facts were used by the student during reasoning. The student also generated a lot of hypotheses with a large number of links (see Scoring method, above). Conversely, the second map of reasoning (bottom of Fig. 1) was almost exclusively inductive (most of the links were drawn in red to show a progression from facts to hypotheses) and was less rich and exhaustive than the first map (richness score: 22; exhaustiveness rate: 0.29). Means and standard deviations of scoring parameters in each group are summarised in Tables 1 and 2.

Reliability

Group W1

Despite some discrepancies in scoring, a positive correlation was found between writers A and B, regardless of the scoring parameters (R coefficients, range 0.77–0.95). The calculated rate of inductive reasoning was positively correlated between writers A and B ($R = 0.89$) with a mean absolute difference between writers of 0.08 (Table 1).

Writer A drew significantly more inductive links (12.3 ± 5.2 versus 10.3 ± 3.1; $p < 0.05$) and hypotheses (9.8 ± 4.6 versus 7.9 ± 3.6; $p < 0.05$) than writer B, which leads to a significant discrepancy in the richness of reasoning score as this parameter was calculated as the sum of links, facts and hypotheses.

Group W2

Correlation coefficients between both writers ranged from 0.66 to 0.93 for each parameter.

Writer D drew fewer inductive links (9.9 ± 4.6 versus 12.3 ± 6.2; $p < 0.05$) and hypotheses (7.1 ± 3.9 versus 9.7 ± 5.9; $p < 0.05$) than writer C. This result leads to a significant difference in the richness of reasoning score (28.4 ± 12.3 versus 37.6 ± 18.8; $p < 0.05$). Writer D also noticed fewer facts than writer C (8.2 ± 3.2 versus 9.5 ± 3.7; $p < 0.05$), which results in a significant difference in the exhaustiveness of reasoning (0.42 ± 0.15 versus 0.49 ± 0.2;

Table 1 Reliability analysis

Scoring parameters in Year 5 medical students (Group 3) generated by the researchers (16 maps per writer)				
	Writer A Mean ± SD	**Writer B Mean ± SD**	**MAD† between writers A and B Mean (min–max)**	**Pearson's or Spearman's coefficient**
Inductive links*	12.3 ± 5.2	10.3 ± 3.1	2.94 (0–8)	0.77
Deductive links	7.9 ± 7.5	6.7 ± 5.9	2.31 (0–6)	0.91
Number of facts	9.7 ± 3.5	9.2 ± 3.7	1.00 (0–3)	0.92
Number of hypotheses*	9.8 ± 4.6	7.9 ± 3.6	2.00 (0–7)	0.89
Rate of inductive reasoning	0.67 ± 0.3	0.67 ± 0.26	0.08 (0–0.35)	0.89
Richness score*	39.8 ± 15	34.2 ± 13	5.75 (1–15)	0.95
Exhaustiveness rate	0.49 ± 0.16	0.46 ± 0.15	0.06 (0–0.25)	0.86
Scoring parameters in Year 4 medical students (Group 2) generated by trained teachers (16 maps per writer)				
	Writer C Mean ± SD	**Writer D Mean ± SD**	**MAD† between writers C and D Mean (min–max)**	**Pearson's or Spearman's coefficient**
Inductive links*	12.3 ± 6.2	9,9 ± 4.6	3.7 (1–8)	0.93
Deductive links	6 ± 6.8	4.3 ± 4.5	3.1 (0–14)	0.78
Number of facts*	9.5 ± 3.7	8.2 ± 3.2	1.9 (0–6)	0.81
Number of hypotheses*	9.7 ± 5.9	7.1 ± 3.9	3.2 (0–12)	0.77
Rate of inductive reasoning	0.71 ± 0.3	0.71 ± 0.3	0.13 (0–0.36)	0.8
Richness score *	37.6 ± 18.8	28.4 ± 12.3	10.1 (0–29)	0.66
Exhaustiveness rate*	0.49 ± 0.2	0.42 ± 0.15	0.11 (0–0.5)	0.66

* $p < 0.05$ (paired Student's or Wilcoxon sign rank tests)
† Mean absolute difference (MAD) estimated by the mean of the absolute value of the differences between writer A and writer B
SD = standard deviation

$p < 0.05$). However, the rate of inductive reasoning remained similar.

Validity

Validity was assessed by looking at variations in three groups scored by the same writer (writer A) in the rate of inductive reasoning according to the type of problem and the participant's level of expertise.

These three groups (Groups 1, 2 and 3) generated 48 maps in total. Rates of inductive reasoning, by case, averaged over all groups were 0.74, 0.68, 0.55 and 0.36 for Problems 1, 2, 3 and 4, respectively and were significantly different ($F = 7.88$, $p < 0.05$). The use of inductive reasoning was significantly lower in Problem 4 than in Problems 1 ($F = 17.78$, $p < 0.05$) and 2 ($F = 14.41$, $p \leq 0.05$), but not Problem 3 ($F = 4.41$, $p < 0.05$; Bonferroni corrections). (Data not shown.)

Significant differences were found in the numbers of inductive and deductive links according to level of expertise (Table 2), which demonstrates a predominantly deductive type of reasoning in Group 3 in comparison with Groups 1 and 2 (rates of inductive reasoning 0.41 versus 0.67 and 0.67, respectively; $P < 0.05$). A decrease of 33% in the number of inductive links was found in the expert group (Group 4) in comparison with the Year 5 student group (Group 3) (8.3 versus 12.3), whereas the number of deductive links increased by 58% in experts compared

Table 2 Comparison of means of scoring parameters in the three groups scored by the same writer (writer A) (n = 16 per group)

	Year 3 (Y3) medical students (Group 1)	Year 5 (Y5) medical students (Group 3)	Specialists (S) in internal medicine (Group 4)	χ^2* or F†	Significant differences‡
Inductive links	9.5 ± 3.0	12.3 ± 5.2	8.3 ± 5.0	3.42† ($p < 0.05$)	Y5 versus S
Deductive links	5.1 ± 3.8	7.9 ± 7.5	12.2 ± 6.3	9.03* ($p < 0.05$)	Y3 versus S
Number of facts	9.4 ± 3.0	9.7 ± 3.5	11.0 ± 3.6	1.04†	
Number of hypotheses	6.7 ± 2.6	9.8 ± 4.6	9.8 ± 5.5	4.88*	
Rate of inductive reasoning	0.67 ± 0.12	0.67 ± 0.30	0.41 ± 0.21	7.39† ($p < 0.05$)	Y3 versus S Y5 versus S
Richness score	30.7 ± 10.9	39.8 ± 15.0	41.3 ± 16.7	2.55†	
Exhaustiveness rate	0.49 ± 0.20	0.49 ± 0.16	0.56 ± 0.18	0.69†	

* Kruskall–Wallis test
† Analysis of variance (ANOVA)
‡ $p < 0.05$; Bonferroni corrections

with Year 3 students (12.2 versus 5.1). Consequently, a decrease of 38% in the rate of inductive reasoning was found in experts compared with both Year 3 and Year 5 students (0.41 versus 0.67, respectively).

No significant difference in the richness of reasoning score was observed between experts and Year 3 students (41.3 ± 16.7 versus 30.7 ± 10.9). No significant difference in the rate of exhaustiveness of reasoning was found between the three groups.

DISCUSSION

This work aimed to explore the feasibility and reliability of a new method to allow rapid identification of the style of clinical reasoning in medical students. This work is the first step towards the building of a new tool for measuring the structure of clinical reasoning. To date and to our knowledge, this method is the first to quantify the respective contributions of inductive and deductive strategies within clinical reasoning. Unlike previous think-aloud research protocols, mapping was performed instantaneously without any prior recording of the verbatim. Asking a third party to write a map in real time represented an initiative that had not previously been used in the field of research in clinical reasoning. Concept mapping by interviewers themselves has been previously carried out, but only for the purpose of creating declarative knowledge learning tools.[32,33]

Taking into account these innovations, it was important to demonstrate the reliability of the method for each scoring parameter, as reported in Table 1. Moreover, even if some differences in terms of the numbers of inductive links and hypotheses were observed, the mean CRIR was similar between writers A and B, with a good level of correlation. The same results were found in writers unfamiliar with the study hypotheses. Moreover, the feasibility of such training is quite workable as it does not require more than 30 minutes.

Validity is a psychometric parameter that is more difficult to assess in a domain that lacks a reference standard. In terms of evidence for construct validity, we found, as expected, a lower CRIR in Problem 4 than in Problems 1 and 2. Contrary to our expectations, Problem 3 did not lead to predominantly inductive reasoning, possibly because of the high level of uncertainty about the correct diagnosis in this case. However, we did not find any difference in the richness and exhaustiveness of the reasoning according to level of expertise, but it is possible that the richness scores suffer from a lack of power. Our results corroborate the data reported by West *et al.*,[16,27] according to which the numbers of concepts and cross-links made by Year 1 and 3 students were very similar. However, our results are discordant with those of Markham and Mintzes,[24] who found a clear increase in concepts, links, cross-links and branching in biology majors compared with non-majors. However, neither of these two

studies concerned the field of clinical reasoning specifically. Interestingly, we found a slight increase in deductive reasoning in experienced doctors (i.e. Group 4), estimated at 59%. This trend towards a predominant type of deductive reasoning in experts must be confirmed in subsequent studies, given the discrepancies in the literature. Indeed, Heemskerk *et al.*[4] found that deductive reasoning was predominant in internal medicine residents (54%), who were slightly younger than our qualified internists, whereas Coderre *et al.*, using a think-aloud protocol,[10,11] showed a clear increase in the use of pattern recognition or scheme-inductive reasoning in experts (94% inductive reasoning in experts versus 41% in novices).

Moreover, the important rates of relative decreases or increases in deductive and inductive links according to level of expertise (leading to large variations in the rate of inductive reasoning) seem to be relevant to the meaningfulness of the statistical differences found in this study.

Our method has two main limitations with opposite effects on our results. Firstly, our use of written simulated problems restricted the use of deductive strategies as all information considered useful was given immediately. This limitation might explain why deductive reasoning was especially poorly represented in students in Years 3 (Group 1) and 5 (Group 3). Secondly, as shown by Eva *et al.*,[34] unconscious reasoning such as pattern recognition, which is considered frequent among experts, may not be recognised with think-aloud protocols. Thus, these authors argued that think-aloud protocols may reflect the diagnostician's explanations of the case better than his or her reasoning approach, which means that we must be cautious about the validity of this method.

In addition, the conclusions of this study are tempered by some considerations of the methodological limits. Firstly, participants were not randomly selected from their class of origin and so may not be representative. Secondly, the fact that map writers were not blinded to the expertise level of participants may have introduced an interpretation bias. Finally, we were not able to calculate the inter-rater correlation between newly trained and expert writers.

This new method for the assessment of problem-solving skills showed a good reliability and some important differences between experts and novices in terms of rates of inductive and deductive reasoning. Although further work is required to understand how to best adapt this method to clinical teaching settings (including the particular challenge of giving feedback), it may prove to be a promising tool in the assessment of medical problem solving. Giving teachers a means of diagnosing how their students think when they are confronted with a clinical problem may facilitate teaching and learning.

Contributors: PP contributed to the conception and design of the study, the acquisition, analysis and interpretation of data, and the write-up of the manuscript. JBH contributed to the conception and design of the study, and the acquisition, analysis and interpretation of data. BDH contributed to the interpretation of data and the revision of the manuscript. BP contributed to the conception and design of the study, the acquisition of data and the revision of the manuscript. MAP contributed to the acquisition and analysis of data, and the revision of the manuscript. JC, RC and CD contributed to the acquisition and analysis of data, and the revision of the manuscript. VS contributed to data acquisition, analysis and interpretation. JHB contributed to the conception and design of the study and the interpretation of data. All authors approved the final manuscript for publication.
Acknowledgements: we would like to acknowledge Dr Nicole Woods at the Wilson Centre, University of Toronto for her valuable editorial revisions and Professor Jean Jouquan for his shrewd advice.
Funding: none.
Conflicts of interest: none.
Ethical approval: in France, no official ethics committee oversees research in medical education. Therefore, before this study began, its protocol was examined and approved by the dean of the Faculty of Medicine, University of Nantes.

REFERENCES

1 Newell A, Simon HA. *Human Problem-Solving*. Englewood Cliffs, NJ: Prentice Hall 1972;1–920.
2 Norman GR, Brooks LR. The non-analytical basis of clinical reasoning. *Adv Health Sci Educ* 1997;**2**:173–84.
3 Norman GR, Young M, Brooks L. Non-analytical models of clinical reasoning: the role of experience. *Med Educ* 2007;**4**:1140–5.
4 Heemskerk L, Norman GR, Chou S, Mintz M, Mandin H, McLaughlin N. The effect of question format and task difficulty on reasoning strategies and diagnostic performance in internal medicine residents. *Adv Health Sci Educ Theory Pract* 2008;**13**:453–62.
5 Patel VL, Arocha JF. Cognitive models of clinical reasoning and conceptual representation. *Methods Inf Med* 1995;**34**:47–56.
6 Arocha JF, Dongwen W, Patel VL. Identifying reasoning strategies in medical decision making: a methodological guide. *J Biomed Inform* 2005;**38**:154–71.

7 Nendaz MR, Bordage G. Promoting diagnostic representation. *Med Educ* 2002;**36**:760–6.
8 Charlin B, Boshuizen PA, Custers EJ, Feltovitch PJ. Scripts and clinical reasoning. *Med Educ* 2007;**41**:1178–84.
9 Fonteyn ME, Grobe SJ. Expert nurses' clinical reasoning under uncertainty: representation, structure, and process. *Proc Ann Symp Comput Appl Med Care* 1992:405–9.
10 Coderre S, Mandin H, Harasym H, Fick GH. Diagnostic reasoning strategies and diagnostic success. *Med Educ* 2003;**37**:695–703.
11 Coderre S, Harasym P, Mandin H, Fick G. The impact of two multiple-choice question formats on the problem-solving strategies used by novices and experts. *BMC Med Educ* 2004;**5**:4–23.
12 Funkesson KH, Anbäcken EM, Ek AC. Nurses' reasoning process during care planning taking pressure ulcer prevention as an example. A think-aloud study. *Int J Nurs Stud* 2007;**44**:1109–19.
13 McLaughlin K, Coderre S, Mortis G, Fick G, Mandin H. Can concept sorting provide a reliable, valid and sensitive measure of medical knowledge structure? *Adv Health Sci Educ Theory Pract* 2007;**12**:265–78.
14 Ruiz-Primo MA, Shavelson RJ. Problems and issues in the use of concepts maps in science assessment. *J Res Sci Teach* 1996;**33**:569–600.
15 McClure JR, Sonak B, Suen HK. Concept map assessment of classroom learning: reliability, validity, and logistical practicability. *J Res Sci Teach* 1999;**36**:475–92.
16 West DC, Park JK, Pomeroy JR, Sandoval JR. Concept mapping assessment in medical education: a comparison of two scoring systems. *Med Educ* 2002;**36**:820–6.
17 McGaghie WC, McGrimmon DR, Mitchell G, Thompson JA, Ravitch MM. Quantitative concept mapping in pulmonary physiology: comparison of student and faculty knowledge structures. *Adv Physiol Educ* 2000;**23**:72–81.
18 McGaghie WC, McGrimmon DR, Mitchell G, Thompson JA. Concept mapping in pulmonary physiology using pathfinder scaling. *Adv Health Sci Educ Theory Pract* 2004;**9**:225–40.
19 Ruiz-Primo MA, Schultz SE, Li M, Shavelson RJ. Comparison of the reliability and validity of scores from two concept-mapping techniques. *J Res Sci Teach* 2001;**38**:260–78.
20 Acton WH, Johnson PL, Goldsmith TA. Structural knowledge assessment: comparison of referent structures. *J Educ Psychol* 1994;**86**:303–11.
21 Novak JD, Gowin DB, Johansen GT. The use of concept mapping and knowledge vee mapping with junior high school science students. *Sci Educ* 1983;**67**:625–45.
22 Stensvold MS, Wilson JT. The interaction of verbal ability with concept mapping in learning from a chemistry laboratory activity. *Sci Educ* 1990;**74**:473–80.
23 Schmid RF, Telaro G. Concept mapping as an instructional strategy for high school biology. *J Educ Res* 1990;**84**:78–85.
24 Markham KM, Mintzes JJ. The concept map as a research and evaluation tool: further evidence of validity. *J Res Sci Teach* 1994;**31**:91–101.
25 Liu X, Hinchey M. The internal consistency of a concept mapping scoring scheme and its effect on prediction validity. *Int J Sci Educ* 1996;**18**:921–37.
26 Hsu L, Hsieh SI. Concept maps as an assessment tool in a nursing course. *J Prof Nurs* 2005;**21**:141–9.
27 West DC, Pomeroy JR, Park JK, Gerstenberger EA, Sandoval J. Critical thinking in graduate medical education. A role for concept mapping assessment? *JAMA* 2000;**284**:1105–10.
28 Brooks LR, Norman GR, Allen SW. Role of specific similarity in a medical diagnostic task. *J Exp Psychol Gen* 1991;**120**:278–87.
29 Kutalunga-Moruzi C, Brooks LR, Norman GR. Coordination of analytic and similarity-based processing strategies and expertise in dermatological diagnosis. *Teach Learn Med* 2001;**13**:110–6.
30 Eva KW. What every teacher needs to know about clinical reasoning. *Med Educ* 2004;**39**:98–106.
31 Mandin H, Jones A, Woloschuck W, Harasym P. Helping students learn to think like experts when solving clinical problems. *Acad Med* 1997;**72**:173–9.
32 Heine-Fry JA, Novak JD. Concept mapping brings long-term movement toward meaningful learning. *Sci Educ* 1990;**74**:461–72.
33 Hoz R, Tomer Y, Tamir P. The relations between disciplinary and pedagogical knowledge and the length of teaching experience of biology and geography teachers. *J Res Sci Teach* 1990;**27**:973–85.
34 Eva KW, Brooks LR, Norman GR. Forward reasoning as a hallmark of expertise in medicine: logical, psychological, and phenomenological inconsistencies. In: Shohov SP, ed. *Advances in Psychological Research*. New York, NY: Nova Scotia 2002;42.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

Appendix S1. Four clinical problems derived from authentic clinical cases and used in this study.

Appendix S2. Instructions for creating and scoring a concept map from student think-aloud processes.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than for missing material) should be directed to the corresponding author for the article.

Received 4 January 2010; editorial comments to authors 18 March 2010, 13 April 2010; accepted for publication 26 April 2010