Reasoning versus knowledge retention and ascertainment throughout a problem-based learning curriculum

Anne Collard,[1,2] Sabine Gelaes,[3] Sophie Vanbelle,[4] Serge Bredart,[5] Jean-Olivier Defraigne,[2] Jacques Boniver[2,3] & Jean-Pierre Bourguignon[1,3]

CONTEXT Since 2000, problem-based learning (PBL) seminars have been introduced into the curriculum of medical studies at the University of Liège. We aimed to carry out a cross-sectional investigation of the maturational increase in biomedical reasoning capacity in comparison with factual knowledge retention throughout the curriculum.

METHODS We administered a factual knowledge test (i.e. a true/false test with ascertainment degree) and a biomedical reasoning test (i.e. an adapted script concordance test [SCT]) to 104 students (Years 3–6) and a reference panel. The selected topic was endocrinology.

RESULTS On the SCT, the students obtained higher scores in Years 5 and 6 than in Years 3 and 4. In Year 3, the scores obtained on SCT questions in a new context indicated transfer of reasoning skills. On the true/false test, the scores of Year 3 students were significantly higher than those of students in the other three year groups. A positive correlation between SCT scores and true/false test scores was observed only for students in Years 3 and 4. In each group, the ascertainment degree scores were higher for correct than for incorrect responses and the difference was calculated as an index of self-estimation of core knowledge. This index was found to be positively correlated to SCT scores in the four year groups studied.

CONCLUSIONS Biomedical reasoning skills are evidenced early in a curriculum involving PBL and further increase during training. This is accompanied by a decrease in factual knowledge retention. The self-estimation of core knowledge appears to be related to reasoning capacity, which suggests there is a link between the two processes.

Medical Education 2009: ***43****: 854–865*
doi:10.1111/j.1365-2923.2009.03410.x

[1]Department of Clinical Sciences, Faculty of Medicine, University of Liège, Liège, Belgium
[2]Department of Biomedical and Pre-clinical Sciences, Faculty of Medicine, University of Liège, Liège, Belgium
[3]Institute of Training and Research in Higher Education (Institut de Formation et de Recherche en Enseignement Supérieur), University of Liège, Liège, Belgium
[4]Department of Public Health Sciences, Faculty of Medicine, University of Liège, Liège, Belgium
[5]Department of Cognitive Sciences, Faculty of Psychology, University of Liège, Liège, Belgium

Correspondence: Jean-Pierre Bourguignon, Department of Clinical Sciences, University of Liège, CHU Sart-Tilman, 13 Avenue de l'Hôpital, 4000 Liège, Belgium. Tel: 00 32 4 366 7247; Fax: 00 32 4 366 7246; E-mail: jpbourguignon@ulg.ac.be

INTRODUCTION

In addition to their necessary learning of a large body of knowledge, medical students must develop clinical reasoning skills. Clinical reasoning has mainly been studied in the context of medical diagnosis, which is considered to involve a problem-solving process. The mental representation of a problem is a key issue in that process. For Feltovich and Barrows,[1] the mental representation of a diagnostic problem takes the form of an (instantiated) 'illness script' into which patient characteristics, signs and symptoms of the disease, and knowledge on the underlying links are organised. Illness scripts may also contain memories of previous patient experiences which help in the diagnosis of new cases.[2] According to Schmidt *et al.*,[2] illness scripts are developed through continuous exposure to patients and thus result from experience. These scripts are used by experienced doctors when diagnosing routine cases. Other authors claim that the first step in the process of clinical reasoning is the early generation of hypotheses, which are then tested and validated through an active process of acquisition of relevant data.[3–7] Although these mechanisms of clinical reasoning are thought to operate mainly by the end of medical studies, when they have been evidenced, we hypothesised that reasoning may be active earlier in the curriculum and may mature subsequently. Findings from studies on clinical reasoning have influenced educational strategies in order to promote reasoning and its assessment. Problem-based learning (PBL) was developed to facilitate the student's process of acquisition, organisation and retrieval of knowledge.[7–10] One aim of PBL is to help the student integrate new information in a rich and connected knowledge network that will be activated later (illness scripts).[11] Through such mechanisms, PBL promotes clinical reasoning skills.[11] Since 2000, the Faculty of Medicine at the University of Liège has integrated PBL seminars into Years 2–7 of a 7-year curriculum.[12] The present study aimed to investigate the maturation of reasoning capacity throughout a medical curriculum.

Some authors have attempted to analyse the relationship between the knowledge base and the maturation of clinical reasoning at different levels of expertise by using a think-aloud protocol applied during clinical problem-solving tasks.[13,14] Patel and Groen[14] have shown that experts use less basic science in their reasoning explanations than novices. Boshuizen and Schmidt[13] demonstrated that the application of biomedical knowledge (use of basic knowledge to explain the process underlying the case) decreased with increasing levels of expertise. They proposed a three-stage model of expertise development in medicine, consisting of the acquisition of biomedical knowledge, practical experience and the integration of both theoretical and experiential knowledge resulting in knowledge encapsulation.[3,15] There are, however, few studies addressing the retention of factual knowledge over time. Some authors have reported decreasing scores for recognition and recall tasks, together with increasing scores for application and interpretation tasks.[16]

In the present study, investigation tools involved a script concordance test (SCT)[17–21] to assess biomedical and clinical reasoning and a true/false test to assess factual knowledge. The maturation of reasoning capacity was compared with knowledge retention. We hypothesised that retention of factual knowledge in the pre-clinical sciences might be highest in Year 3 and might decrease afterwards, whereas reasoning capacity might be lowest in Year 3 and might further increase throughout the curriculum. Degree of ascertainment is used as part of routine procedure for multiple-choice question and true/false evaluations in our institution.[22–24] The ascertainment degree measures how sure the responder is about each of his or her answers and, thus, gives an indication of the subject's self-estimation of what he or she knows and does not know.[22–24] The relationship between self-estimation of core knowledge (through the ascertainment degree on the true/false test) and reasoning was investigated.

METHODS

Participants

During the academic year 2006–2007, 104 volunteer students from Years 3–6 in the Faculty of Medicine at the University of Liège participated in the study (a minimum of 20 students in each study year). Endocrinology was the selected topic of the investigations. The students differed in terms of the PBL seminars and clerkship periods they had undertaken (Table 1). In Years 2 and 3, the majority of learning occurs in PBL seminars (named APP, Apprentissage par Problèmes),[12,25] which focus on the structure and functioning of the human body systems. In Years 4 and 5, the diagnostic and therapeutic bases of diseases of the different systems are learned through lectures and PBL seminars (named ARC, Apprentissage du Raisonnement Clinique).[12,25] There was no difference among the groups of volunteer students in terms of their global academic achievement (assessed on the basis of Year 3 notes). Likewise, volunteers did not differ from

Table 1 Participant characteristics

	Year 3		Year 4		Year 5		Year 6	
	Volunteers	Non-participants	Volunteers	Non-participants	Volunteers	Non-participants	Volunteers	Non-participants
n	35	79	20	77	25	71	24	68
Gender, F/M	24/11	51/28	10/10	55/22	13/12	46/25	13/11	33/55
Mean score, Year 3	13.9	13.4	14.4	13.7	14.5*	13.3	14.1	13.5
Completed seminars (endocrine system) at the time of study	APP		APP		APP and ARC		APP and ARC	
Lectures on diagnosis and treatment of endocrine disorders	–		–		Done		Done	
Completed clerkship (all disciplines) at the time of study	–		Half-time 4 months		Half-time 12 months		Half-time 12 months and full-time 4 months	
Reference panel	9 tutors involved in APP and ARC seminars about the endocrine system							

* $P < 0.05$ versus non-participants Year 5.
F = female; M = male; APP = Apprentissage par Problèmes (problem-based learning); ARC = Apprentissage du Raisonnement Clinique (clinical reasoning).

non-participants, except for in Year 5, in which volunteer notes were of significantly better quality than those of the rest of the Year 5 cohort. Written informed consent was obtained from the participants and tests were completed anonymously. Confidentiality and anonymity were guaranteed. No student names were used in the statistical analysis. Each participant was assigned a number with which he or she was able to check his or her scores online and compare them with the whole-group scores.

Two teachers wrote the testing materials. Nine tutors (three in academic appointments and six in clinical appointments at the university hospital), who handled the PBL seminars on the endocrine system, served as a reference panel for SCT scoring (see below). They also provided answers to the true/false test. Written informed consent was obtained from the tutors, who completed the tests anonymously.

Test materials

Clinical reasoning test: the script concordance test

Originally, the SCT[17–21] proposed a series of short and incomplete clinical scenarios, each of which was followed by either a diagnostic hypothesis or a proposal for the therapeutic management of the patient. When new information was provided, the responders had to assess the impact of this information on the basis of the initial diagnostic or management hypothesis. Responses were given on a 5-point Likert scale. Because the incomplete clinical scenario allowed different answer options, the response was rarely unequivocal. Each individual response was credited by a score determined on the basis of the distribution of responses given by a reference panel.[18,19,26]

Starting from the initial SCT format, we adapted the test to assess skills of biomedical reasoning in students not yet familiar with the process of clinical diagnosis. The SCT had not previously been used for this purpose. The present study was based on 48 SCT questions with scenarios inspired by the clinical vignettes used in the PBL seminars (APP and ARC). The SCT scenarios were based on clinical cases taken from PBL seminars on endocrinology: five cases were sourced from Year 3 seminars (APP) and five cases from Year 4 seminars (ARC). For each clinical case, two SCT scenarios which aimed at the same learning objectives were created, either in the original learning context or in a new context (in which case characteristics differed from those of the learning clinical cases). For each scenario, three independent hypotheses were proposed, each followed by newly given information. A single hypothesis could be used three times, as in the example provided in Table 2. The task was to evaluate the impact of new information on the initial hypothesis. From a total of 60 administered questions, 12 were discarded by the reference panel because the scenario was misunderstood as a result of unclear writing.

Factual knowledge test: true/false test

A 120-question true/false test was elaborated. No clinical scenario was presented in vignette in this test format. The questions explored factual knowledge through statements. They were related to the same topics as the SCT, which enabled the comparison between the maturation of biomedical reasoning and the retention of factual knowledge during medical studies. Sample questions are provided in Table 3. For each true/false test answer, an ascertainment degree[22–24] was requested to measure how sure the participant was of his or her response. The ascertainment degree was given on a 6-point Likert scale (Table 3) ranging from 0 (the response has a probability of 0–25% of being correct) to 5 (the response has a probability of 95–100% of being correct). This validated scale[22–24] has been used for several years in the certificating evaluation at the Faculty of Medicine at Liège. For each responder, the mean ascertainment degree was calculated separately for correct answers versus incorrect answers. The difference between the mean ascertainment degrees (correct–incorrect answers) was used as an estimate of the capacity for self-discrimination between known and unknown material.

Test administration

The tests were administered at the beginning of the second quarter of the academic year (February 2007). All the students were gathered to complete the tests at the same time. The reference panel completed the tests before the students. Each student and tutor started with the SCT, which was followed by the true/false test. There was no time limit. The tests were completed individually. We asked the participants to answer all questions on both tests. We created three different versions of each test in which the questions appeared in a pseudo-random order.

Table 2 Example of a script concordance test question and referee panel answers. A 50-year-old pre-menopausal woman shows up for a routine visit in the Department of Occupational Medicine. Her body mass index is 28; she is sedentary. Glycosuria is found at screening urine analysis

If the hypothesis is	And you know that	The hypothesis becomes	Reference panel answer
Glycosuria results from early-stage type 2 diabetes	The patient takes medication for hypercholesterolaemia	Much less likely	
		Less likely	3
		Not affected	6
		More likely	
		Much more likely	
Glycosuria results from early-stage type 2 diabetes	Together with a fasting blood sugar of 126 mg/dl, plasma insulin level was high	Much less likely	
		Less likely	
		Not affected	
		More likely	2
		Much more likely	7
Glycosuria results from early-stage type 2 diabetes	She was fasting at the time of urine collection	Much less likely	
		Less likely	1
		Not affected	1
		More likely	3
		Much more likely	4

Table 3 Examples of true/false (T/F) test questions

Questions	T/F	Ascertainment degree
1 In the adipose tissue, insulin activates fatty acid synthesis		0 (0%–25%) | 1 (25%–50%) | 2 (50%–70%) | 3 (70%–85%) | 4 (85%–95%) | 5 (95%–100%)
2 Insulin causes an increase in plasma-free fatty acid concentrations		0 (0%–25%) | 1 (25%–50%) | 2 (50%–70%) | 3 (70%–85%) | 4 (85%–95%) | 5 (95%–100%)
3 The insulin receptor has enhanced affinity for insulin in type 2 diabetes		0 (0%–25%) | 1 (25%–50%) | 2 (50%–70%) | 3 (70%–85%) | 4 (85%–95%) | 5 (95%–100%)

Scoring process on the SCT

Responses to SCT questions were marked using a modified scoring method.[27] The original scoring process used by Charlin and collaborators was 'the aggregate scoring process'.[18,21] The score attributed to the responder for each item of a question was based on the proportion of panel members who had selected the same answer. The maximum score was 1 for the modal answer. The number of panel members who provided an answer on the Likert scale was divided by the number of experts in the modal value for the item. To give an example, we refer to the referee responses to the sample question illustrated in Table 2. On the last item, the highest number (four of nine) of referees responded with 'much more likely' on the Likert scale. This option was thus credited with the maximum score of 1 point (4/4 [i.e. the modal response]). Then, as three referees had selected 'more likely' as a response, this option was credited with 0.75 points (3/4). The answers 'not affected' and 'less likely' were credited with 0.25 points (1/4) each. The answer(s) not selected by any of the panel members ('much less likely' in this example) received no credit, whether they were close to or far from the modal answer. Here, the scoring method used by Charlin *et al.* was adapted using the kappa coefficients methodology[27] to account for scores obtained fortuitously. A weighted kappa coefficient[28] quantified the agreement between two raters on an ordinal scale and corrected for agreement resulting from chance. This methodology was extended to quantify the agreement between the reference panel and a responder. The weights permitted higher scores to be assigned to responses closer to the modal response(s) given by the reference panel. The weights were derived from the reference panel responses by computing the frequency distribution of the number of category differences separating two experts' responses for all possible pairs of experts. This led to a weighting of 1.00 for a category of the Likert scale consistent with the modal response of the reference panel, a weighting of 0.96 for the adjacent categories and weightings of 0.19, 0.06 and 0.00 at distances of two, three and four categories, respectively. Cohen's kappa coefficient for the reference panel was also calculated.

Statistical analysis

The results on the SCT and true/false tests were expressed as mean (± standard deviation [SD]). On the SCT, the effect of experience (curriculum level) of the responders was tested using a one-way ANOVA. On the true/false test, Cronbach's alpha was used to assess reliability and performances throughout the curriculum were analysed using a one-way ANOVA. The ascertainment degree was analysed using a repeated-measures ANOVA with the correctness of answer as a within-subjects factor.

Subsequent a priori planned comparisons were used to determine significant differences. Pearson's correlations were used to characterise the association between SCT and true/false test scores and between SCT and the self-discrimination of knowledge (estimated by the difference between ascertainment degrees for correct and incorrect answers on the true/false test). Missing data were kept as such. The results were considered to be significant at the 5% probability level ($P < 0.05$).

RESULTS

SCT scores

The mean scores of the five groups are given in Table 4. The mean scores differed significantly among the five groups ($F_{[4,108]} = 12.56$, $P < 0.0001$). More specifically, the reference panel showed

Table 4 Script concordance test (SCT) score, true/false test score and ascertainment degree in relation to level of study in a medical curriculum

			True/false test					
					Ascertainment degree (%)			
	SCT score		Score		Correct answers		Incorrect answers	
Group (*n*)	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Year 3 (35)	0.57	0.16	75.8*	8.22	71.1†	11.9	60.3	14.9
Year 4 (20)	0.55	0.15	65.6	6.16	57.8†	15.1	49.8	13.7
Year 5 (25)	0.67‡	0.10	68.6	3.90	68.2†	10.1	58.4	11.3
Year 6 (24)	0.64‡	0.11	66.3	6.29	58.2†	12.6	48.6	13.4
Tutors (9)	0.79§	0.05	72.6	7.46	68.7†	15.6	55.1	17.1

Script concordance test scores are Cohen's kappa scoring the degree of agreement between the individual responder and the whole panel of tutors in the different groups. They vary between − 1 and + 1.
True/false data included the test score and the ascertainment degree for correct and incorrect answers. The true/false test scores vary between 0 and 100.
* $P < 0.05$ versus Years 4–6.
† $P < 0.05$ versus ascertainment degree for incorrect answers.
‡ $P < 0.05$ versus Years 3 and 4.
§ $P < 0.05$ versus Years 3–6.
SD = standard deviation.

higher scores than all groups of students; students in Years 5 and 6 obtained higher scores than students in earlier years (Years 3 and 4), with a reduced scattering of the data as indicated by the SDs. As shown in Fig. 1, the maturational increase in SCT scores was observed similarly using questions related to PBL seminars run in Year 3 (APP) and in Years 4 and 5 (ARC). It was noteworthy that students in Years 3 and 4 performed better on ARC than APP questions, although the former had not yet attended any ARC seminars and the latter had not yet attended endocrinology ARC seminars.

True/false test scores and ascertainment degree

Cronbach's alpha coefficient was 0.79 for the true/false test, indicating acceptable reliability. The mean scores of the five groups are given in Table 4 and shown in Fig. 1. Results of the true/false test differed among the groups ($F_{[4,108]} = 10.54$, $P < 0.0001$). The scores of Year 3 students were significantly higher ($P < 0.05$) than the scores obtained by all other student groups (Years 4–6). Moreover, the scores of the reference panel did not significantly differ from those of Year 3 students, but were higher than those of students in Years 4 and 6.

Ascertainment degree analysis showed significantly higher scores for correct than for incorrect responses ($F_{[1,108]} = 224.11$, $P < 0.0001$), whatever the group of participants. The individual differences in ascertainment degree between correct and incorrect answers are shown in Fig. 2. The data show similar values and distributions in the four student groups, as well as among the tutors. No significant correlation was found in any group between ascertainment degree for correct answers and the difference between ascertainment degrees for correct and incorrect answers (data not shown).

SCT scores versus true/false test scores and ascertainment degree

Although the SCT scores increased during the medical education curriculum, the true/false test scores decreased after Year 3 (Table 4, Fig. 1). Because the differences in SCT scores were not significant between Years 3 and 4, and Years 5 and 6, respectively, these groups were combined for the

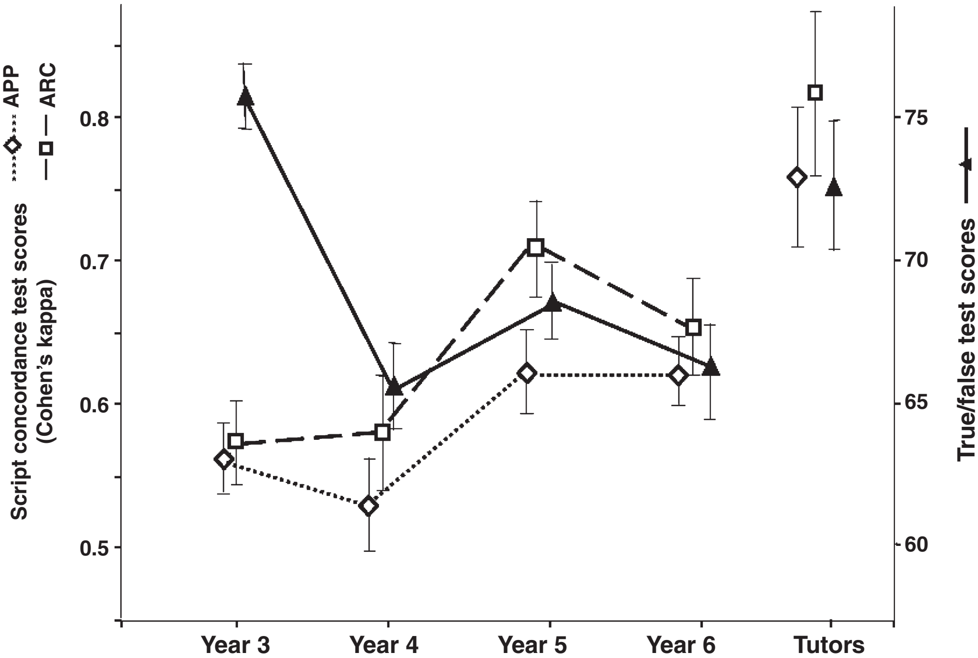

Figure 1 Changes in reasoning skills (evaluated using Cohen's kappa for the script concordance test [SCT]) and knowledge retention (evaluated through scores on the true/false test) occurring throughout a medical curriculum (Years 3–6) in comparison with those of a panel of tutors. The SCT data are shown separately for questions related to problem-based learning (Apprentissage par Problèmes [APP]) seminars (run in Year 3) and clinical reasoning (Apprentissage du Raisonnement Clinique [ARC]) seminars (run in Years 4 and 5)

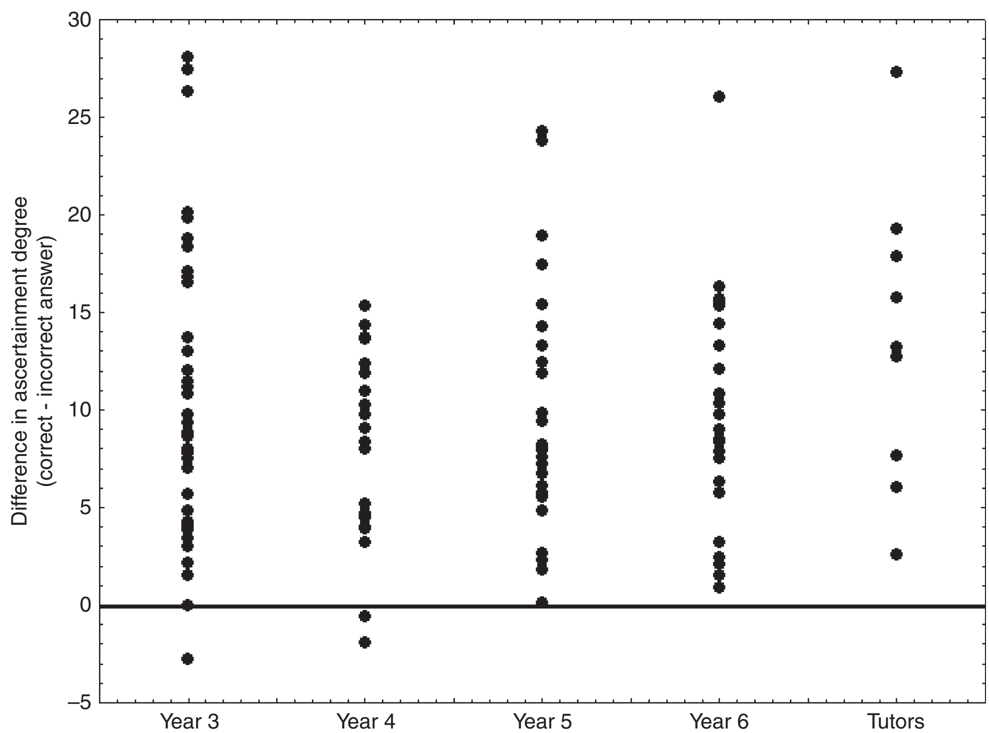

Figure 2 Individual differences between ascertainment degrees provided for correct and incorrect answers at four stages in the medical curriculum and in tutors

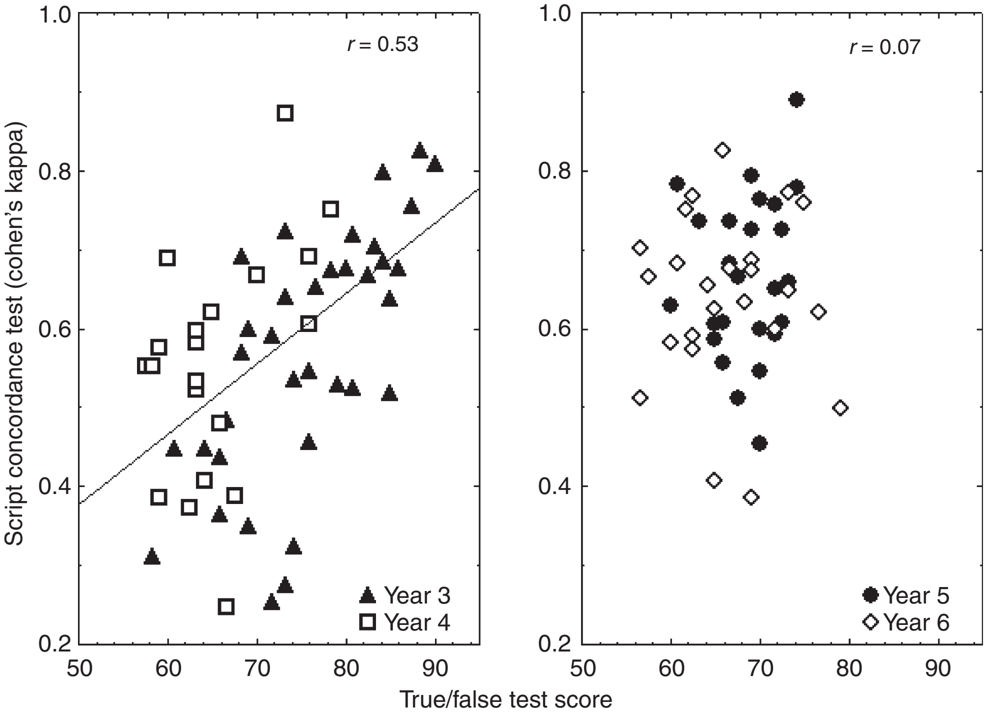

Figure 3 Comparisons of correlations between scores on the true/false test and the script concordance test in two groups of medical students at different stages in the curriculum

correlation analysis. A significant positive correlation (Fig. 3) between SCT scores and true/false test scores was observed for students in Years 3 and 4 ($r = 0.53$, $P < 0.0001$), but not for those in Years 5 and 6 ($r = 0.07$, $P = 0.64$). A significant positive correlation between SCT scores and differences in ascertainment degrees for correct and incorrect answers (Fig. 4) was observed for all students (Years 3 and 4: $r = 0.37$, $P = 0.006$; Years 5 and 6: $r = 0.32$, $P = 0.03$).

DISCUSSION

It appears from the present study that the SCT can be adapted to assess biomedical and clinical reasoning at different stages in a medical curriculum. There are several validation issues that must be addressed. Validity of the content refers to consistency with teaching content, which was ensured by involving two teachers in the writing of the test. Validation for clarity of writing would have required pre-validation of the SCT questions with tutors and students. This, however, would have resulted in the exclusion of those test persons from an already limited group of participants. Therefore, we elected to first administer the whole test to both the students and tutors and then, in a second process, to review the questions for clarity and consistency with the panel of tutors. Following this process, 12 of the 60 questions were invalidated and were not included in the analysis. The significance of our adapted SCT for assessment of reasoning skills was unfortunately not validated directly. We thought about making a possible validation through correlating the SCT scores with notes obtained at a certificating examination which involved a pathophysiological discussion of a short clinical scenario. However, this was incompatible with the anonymous and non-certificating nature of our study.

Although this is the first time that SCT scores have been shown to increase in the course of a PBL medical education, our findings are consistent with observations made later during training, which show that SCT scores of expert doctors are increased over those of postgraduate students.[17–21] It is of note that the reasoning skills of Year 3 students were already obvious and could be applied in the context of ARC-related scenarios to which those students had not been previously exposed. Such a finding suggests that reasoning skills can be transferred to a context that differs from that in which they were learned. The SCT results show a significant maturational increase in terms of biomedical and clinical reasoning capacity between the two earlier and two later year groups studied. These outcomes confirm previous findings

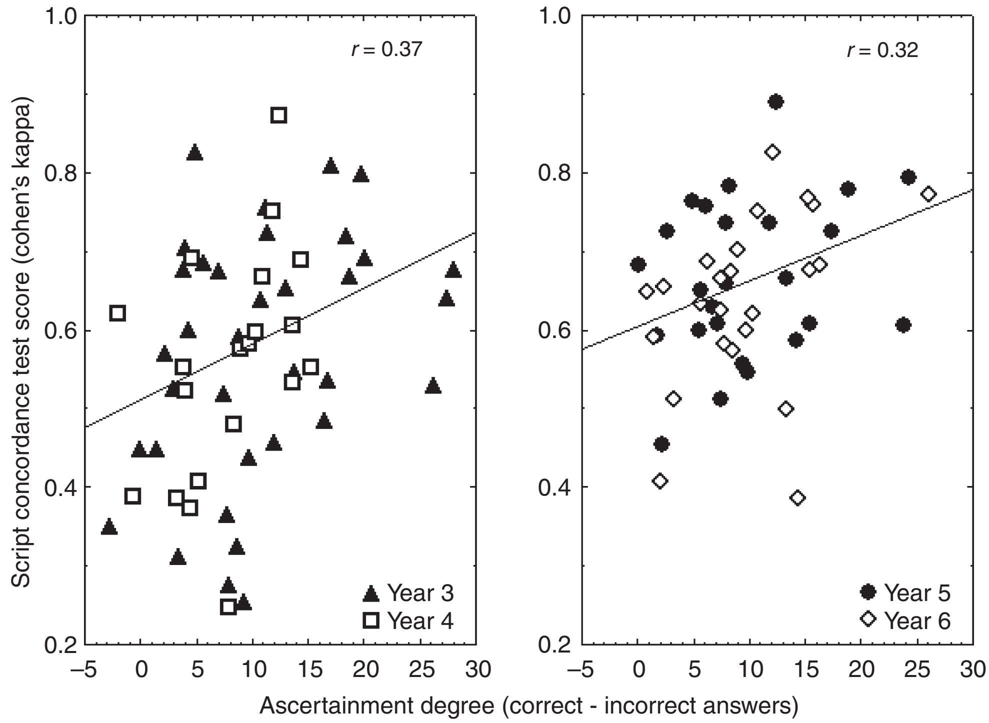

Figure 4 Correlations between script concordance test scores and differences between ascertainment degrees for correct and incorrect answers on the true/false test compared in two groups of medical students at different stages in the curriculum

showing that gaining experience and being exposed to patients contributes to the development of clinical reasoning.[2] This is consistent with the substantial increase in the time periods our students spend in clerkships during Years 5 and 6 in comparison with Years 3 and 4. However, evaluation of the possible contributions of PBL seminars or clerkships or any other learning activity in this maturation process was not the aim of the present study. Because our training curriculum uses a mixed system which combines PBL seminars and other modalities, including formal lectures, it would be almost impossible to establish whether the skills observed on the SCT are related to a specific part of or change in training. Rather, it is fair to state that the introduction of PBL seminars into the curriculum was a reason for teachers and students to promote reasoning skills and for us, as investigators, to attempt an evaluation of those skills.

A strength of this study concerns the original application of a single reasoning test throughout the medical curriculum in a cross-sectional sample of students, all of whom experienced similar learning conditions, including PBL seminars. A longitudinal prospective study would carry some added value as it would result in increased homogeneity and comparability within the student sample. However, in such conditions, different tests bearing on different topics would be administered, thus diminishing their comparability. In addition, enhancements in student capacity as the curriculum progresses, which reflect students' increasing familiarity with a new testing procedure, may bias a longitudinal study. Finally, the number of student volunteers may decrease with time and invalidate the study through an insufficient number of participants.

Results on the true/false test revealed that student scores decreased in Years 4–6 of the curriculum, compared with those of Year 3 students, which were not significantly different from those of the reference panel. The higher scores obtained by Year 3 students may be explained by the fact that these students had just sat an examination on the topic involved in the test and were still easily able to retrieve the knowledge acquired from their memory. For students in the later year groups, this knowledge might be more difficult to retrieve because a longer period had elapsed since it had been learned and other unrelated knowledge had been processed in the interim. Moreover, organisation of the knowledge network may mature less rapidly than acquisition of knowledge, which would account for a lower yield of reactivation, particularly in a test in which contextualised information is not triggered. The knowledge networks of the reference panel may have been better organised and more easily activated because the panel consisted of specialists in the study domain who processed the information in their daily practice. Expert involvement in the clinical practice of endocrinology was more likely to account for the higher scores on the true/false test than global expertise independent of the study domain. These observations are consistent with the data reported by Boshuizen and Schmidt[13] and with their encapsulation theory, which proposes that basic knowledge is still available even if experts do not process basic knowledge in itself when they make a diagnosis.[13,15,29,30] In addition, some members of the reference panel taught endocrinology and therefore regularly reactivated and manipulated knowledge in this particular domain. Presumably, the true/false test study of performances of expert doctors in a domain other than endocrinology would result in a performance level similar to or lower than that of Year 6 students.

The increasing scores on the biomedical reasoning test and the decreasing scores on the factual knowledge test throughout the medical curriculum were in line with results of studies comparing performance in remembering tasks and application tasks.[16] In some studies evaluating the maturation of clinical reasoning and knowledge, both reasoning and knowledge were found to increase during the curriculum. In this particular case, however, knowledge was evaluated in a contextualised diagnostic processing[31,32] (based on clinical scenarios). By contrast, the retention of knowledge acquired in the first years of training was studied here through non-contextualised questions. Thus, the discrepancy between the two studies might suggest the facilitating impact of context on knowledge retrieval. This hypothesis could be substantiated in a study addressing this issue specifically, bearing in mind that the effects of context on knowledge retrieval from the memory network should be differentiated from the possible role played by context at the time of knowledge acquisition (i.e. during a learning process, such as PBL seminars).

A positive correlation between scores on the SCT and the true/false test was observed for students in the earlier years (Years 3 and 4), but not for those in the later years. Such correlations suggest that, in students at the beginning of a PBL curriculum, clinical reasoning skills are directly linked with level of background knowledge. The absence of any significant correlation in students in the later years may indicate that a relative independence of factual knowledge and clinical reasoning has developed with experience. Throughout maturation, reasoning may become more related to experience and matching with known cases, whereas knowledge is not influenced in the same manner. According to Patel *et al.*,[33] expert performance is not a result of generally superior memory skills. It is, rather, a function of a well-organised knowledge base equipped to recognise a familiar configuration of stimuli. This superior quality of experts' mental representations allows them to rapidly select and use relevant information, and to adapt the processing of information to changing circumstances. As practitioners gain experience in executing a task, their performance becomes increasingly smooth, efficient and automatic.[31] The difference between expert and less skilled individuals is not merely a matter of the amount and complexity of accumulated knowledge; it mostly reflects qualitative differences in cognitive strategies, knowledge organisation and problem representations.[33] This concept is consistent with our finding that reasoning capacity increases throughout the medical curriculum. It is, however, interesting to observe that Year 3 medical students show obvious reasoning skills. Further studies are warranted to delineate the possible determinants of such early development of reasoning capacity, especially the possible contribution of PBL seminars.

The difference between ascertainment degrees provided for correct and incorrect answers on the true/false test has been used here to indicate the capacity of self-estimation of core knowledge on the part of the participants. It was anticipated that, when maximal, such a capacity would result in degrees of ascertainment that were high for correct answers and low for incorrect answers, and that this would be translated into a marked difference between the two degrees. By contrast, a minimal self-estimation capacity would be translated into a small difference. This capacity, however, was not related to experience as there were no differences among the four groups of students and the tutors. Furthermore, the capacity to discriminate correct and incorrect answers was not related to the global level of ascertainment for correct answers. A positive correlation between SCT scores and the differences in ascertainment for correct and incorrect answers on the true/false test was observed in all student groups. These results suggest that either the ability to discriminate what is known and what is unknown has common determinants with reasoning capacity, or that the two parameters can influence each other. The higher the difference between the ascertainment degrees provided for correct and incorrect answers, the better the reasoning capacity. This finding is crucial because the effectiveness of reasoning depends on the correctness of the knowledge retrieved during the reasoning process. Such an issue is worthy of further study, especially in a PBL curriculum where reasoning capacity is stimulated early in the training.

In summary, in a 7-year medical curriculum in which PBL seminars are integrated with other learning modalities, Year 3 students demonstrated acquired reasoning skills (assessed through an adapted SCT) that they were able to transfer into a new context. These skills increased throughout the curriculum, whereas knowledge retention showed a rapid reduction. A direct correlation between knowledge and reasoning skills was observed initially and disappeared with maturation. A new finding, however, showed that student capacity to self-estimate core knowledge was directly related to reasoning skills in the four year groups studied. In further research, it would be interesting to compare these results with those of other tests assessing the same processes of knowledge retention and biomedical reasoning that we studied here. In addition, the relationship between reasoning, and knowledge and its self-assessment, should be investigated in more depth. From an educational perspective, the extent to which the individual student is stimulated to develop reasoning skills and to validate his or her core knowledge in the learning activities appears to be important. It may be that greater emphasis is more commonly placed on the former aspect in PBL and on the latter in formal lectures. In further research, we will investigate the possible impact of activities integrated into APP seminars and directed towards self-estimation of core knowledge.

Contributors: AC, J-PB and SG acted as principal investigators on this project and contributed towards the creation of test materials, collection, analysis and interpretation of data, and the writing of the manuscript. SV selected the statistical methods, particularly for the script concordance test, performed the analysis, wrote the statistical section and contributed to the revision of the manuscript. SB served as a co-promoter of this project and contributed towards data interpretation, and the review and editing of the manuscript. J-OD was involved in creating test materials and the revision of the manuscript. JB supported the project in the Faculty of Medicine and contributed towards data interpretation and the editing of the manuscript. All authors approved the final manuscript for publication.
Acknowledgements: the authors are indebted to the tutors who participated in the reference panel (V Geenen, E Hamoir, S Maweja, N Paquot, A Pintiaux, M Rorive, H Valdes, J-P Squifflet, A-S Parent and M Meurisse), to Professors A Albert for the supervision of statistical analysis, and B Charlin and D Leclercq for recommendations, and to the medical students who kindly volunteered to participate in this study.
Funding: this research project was supported by the Belgian Fonds de la Recherche Fondamentale Collective (grants G.4571.06 and 2.4593.08).
Conflicts of interest: none.
Ethical approval: this study was approved by the Belgian National Research Foundation, Bruxelles.

REFERENCES

1 Feltovich PJ, Barrows HS. Issues of generality in medical problem solving. In: Schmidt HG, De Volder ML, eds. *Tutorials in Problem-based Learning: a New Direction in Teaching the Health Professions*. Assen: Van Gorcum 1984;128–42.
2 Schmidt HG, Norman GR, Boshuizen HPA. A cognitive perspective on medical expertise: theory and implication. *Acad Med* 1990;**65**:611–21.
3 Schmidt HG, Boshuizen HPA. On acquiring expertise in medicine. *Educ Psychol Rev* 1993;**5**:205–21.
4 Elstein AS. What goes around comes around: return of hypothetico-deductive strategy. *Teach Learn Med* 1994;**6**:121–3.
5 Barrows HS, Norman GR, Neufeld VR, Feightner JW. The clinical reasoning of randomly selected physicians in general medical practice. *Clin Invest Med* 1982;**5**:49–55.
6 Bordage G. Elaborated knowledge: a key to successful diagnostic thinking. *Acad Med* 1994;**69**:883–5.
7 van Gessel E, Nendaz MR, Vermeulen B, Junod A, Vu NV. Development of clinical reasoning from the basic sciences to the clerkships: a longitudinal assessment of medical students' needs and self-perception after a transitional learning unit. *Med Educ* 2003;**37**:966–74.
8 Schmidt HG. Problem-based learning: rationale and description. *Med Educ* 1983;**17**:11–6.
9 Norman GR, Schmidt HG. The psychological basis of problem-based learning: a review of the evidence. *Acad Med* 1992;**67**:557–65.
10 Regehr G, Norman GR. Issues in cognitive psychology: implications for professional education. *Acad Med* 1996;**71**:988–1001.
11 Barrows H. The scope of clinical education. *J Med Educ* 1986;**61**:23–33.
12 Boniver J. Les études de médecine à l'Université de Liège: le renouveau pédagogique de la Faculté de Médecine. *Rev Med Liege* 2004;**59** (12):717–30.
13 Boshuizen HPA, Schmidt HG. On the role of biomedical knowledge in clinical reasoning by experts, intermediates and novices. *Cogn Sci* 1992;**16**:153–84.
14 Patel VL, Groen GJ. Knowledge-based solution strategies in medical reasoning. *Cogn Sci* 1986;**10**:91–110.
15 Schmidt HG, Rikers R. How expertise develops in medicine: knowledge encapsulation and illness script formation. *Med Educ* 2007;**41**:1133–9.

16 Semb GB, Ellis JA. Knowledge taught in school: what is remembered? *Rev Educ Res* 1994;**64** (2):253–86.
17 Caire F, Sol JC, Charlin B, Isodori P, Moreau JJ. Le test de concordance de script (TCS) comme outil d'évaluation formative des internes en neurochirurgie: implantation du test sur Internet à l'échelle nationale. *Pédagogie Médicale* 2004;**5**:87–94.
18 Charlin B, Gagnon R, Sibert R, van der Vleuten C. Le test de concordance de script, un instrument d'évaluation du raisonnement clinique. *Pédagogie Médicale* 2002;**3**:135–44.
19 Charlin B, Roy L, Brailovsky C, Goulet F, van der Vleuten C. The script concordance test, a tool to assess the reflective clinician. *Teach Learn Med* 2000;**12**:189–95.
20 Charlin B, St-Jean M. Le test de concordance de script: un outil pour évaluer le jugement en médecine. *Bulletin du CEFES, Université de Montréal* 2002;**6**:4–5.
21 Charlin B, van der Vleuten C. Standardised assessment in the context of uncertainty: the script concordance approach. *Eval Health Prof* 2004;**27**:304–19.
22 Leclercq D. *L'évaluation subjective de la probabilité d'exactitude des réponses en situation pédagogique.* [Doctorate thesis in Education Science, University of Liège.] 1975.
23 Leclercq D. Confidence marking, its use in testing. In: Choppin B, Postlethwaite N, eds. *Evaluation in Education.* Oxford: Pergamon Press 1982;161–287.
24 Leclercq D, Boxus E, de Brogniez P, Lambert F, Wuidar H. The TASTE Approach: General implicit solutions in MCQ, open books exams and interactive testing and self-assessment. In: Leclercq D, Bruno J, eds. *Item Banking, Interactive Testing and Self-Assessment.* Berlin: Springer Verlag 1993;210–32.
25 Barrows HS, Tamblyn R, eds. *Problem-based Learning: an Approach to Medical Education.* New York, NY: Springer 1980.
26 Charlin B, Tardif J, Boshuizen P. Scripts and medical diagnosis knowledge: theory and applications for clinical reasoning instruction and research. *Acad Med* 2000;**75**:182–90.
27 Vanbelle S, Massart V, Giet D, Albert A. Test de concordance de script: un nouveau mode d'établissement des scores limitant l'effet du hasard. *Pédagogie Médicale* 2007;**8**:71–80.
28 Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement of partial credit. *Psychol Bull* 1968;**70**:213–20.
29 Rikers R, Loyens S, te Winkel W, Schmidt HG, Sins PHM. The role of biomedical knowledge in clinical reasoning: a lexical decision study. *Acad Med* 2005;**80** (10):945–9.
30 Rikers R, Schmidt HG, Moulaert V. Biomedical knowledge: encapsulated or two worlds apart? *Appl Clin Psychol* 2005;**19** (2):223–31.
31 Boshuizen HPA, van der Vleuten C, Schmidt HG, Machiels-Bongaerts M. Measuring knowledge and clinical reasoning skills in a problem-based curriculum. *Med Educ* 1997;**31**:115–21.
32 Verhoeven B, Verwijnen G, Scherpbier A, van der Vleuten C. Growth of medical knowledge. *Med Educ* 2002;**36**:711–7.
33 Patel VL, Glaser R, Arocha JF. Cognition and expertise: acquisition of medical competence. *Clin Invest Med* 2000;**23** (4):256–60.

Received 22 May 2008; editorial comments to authors 6 August 2008, 4 March 2009; accepted for publication 30 April 2009