Instr Sci (2014) 42:579–594
DOI 10.1007/s11251-013-9290-5

Do expert clinical teachers have a shared understanding of what constitutes a competent reasoning performance in case-based teaching?

Geneviève Gauthier · Susanne P. Lajoie

Received: 14 February 2013 / Accepted: 17 August 2013 / Published online: 19 September 2013

Abstract To explore the assessment challenge related to case based learning we study how experienced clinical teachers—i.e., those who regularly teach and assess case-based learning—conceptualize the notion of competent reasoning performance for specific teaching cases. Through an in-depth qualitative case study of five expert teachers, we investigate whether they share a common concept of what constitutes a good reasoning performance for a set of three teaching cases. We ask expert teachers to reflect on their problem-solving performances to extract specific expectations regarding the assessment of learners. Using visual representations of their performance, experts inspect and identify whether key elements are considered critical, necessary, and useful for the assessment of learners' performance. Findings indicate that despite solving cases differently, expert teachers share a common concept regarding the key elements that demonstrate good clinical reasoning for specific cases. These results and methods used to trigger assessment criteria from expert clinical teachers show potential for the development of process measures in the assessment of clinical reasoning.

Keywords Assessment of problem solving · Teachers' judgment · Pedagogical content knowledge · Clinical reasoning · Case-based learning · Computer-based learning environment

Case-based learning approaches in professional education use teaching cases to provide learners with opportunities to engage in complex problem solving and reasoning tasks (Merseth 1994; Schön 1987; Shulman 1992). In medical education, for example, a teaching case typically consists of a story about one or more problems affecting a

G. Gauthier (✉)
Department of Educational Psychology, University of Alberta, 6-102 Education North, Edmonton, AB T6G 2G5, Canada
e-mail: genevieve.gauthier@ualberta.ca

S. P. Lajoie
Department of Educational Psychology, McGill University, Montreal, QC, Canada

patient and a scenario for addressing the diagnosis and management of these problems (Barrows and Tamblyn 1980; Hmelo 1998). Solving these cases requires that learners engage with a problem prior to receiving feedback and debriefing about how to resolve the problem in practice. Despite well researched benefits related to these teaching approaches (Albanese and Mitchell 1993; Dochy et al. 2003; Williams 1992), the assessment of case-based learning is problematic (Savin-Baden, 2004). The learning that occurs in small-scale interactive case-based contexts is not consistent with the underlying learning framework of current assessment tools based on high-stakes testing, which are difficult to implement in small-scale interactive contexts (Brookhart 2003; Shepard 2006).

In professional education contexts, the use of teaching cases is gaining popularity; learners are requesting more cases and accreditation program reviews are promoting the integration of teaching cases into the formal curriculum. As a result, many institutions are developing technology-mediated support to promote and oversee the use of teaching cases in their curricula (Cook 2007; Hmelo-Silver et al. 2010). Adapting these case-teaching activities anchored in an oral narrative tradition (Greenhalgh and Hurwitz 1999) into a technology-mediated format presents opportunities to study teachers' assessment practices in case-based learning contexts.

We propose to explore the assessment challenge by studying how experienced clinical teachers—i.e., those who regularly teach and assess case-based learning—conceptualize the notion of a good answer for specific cases through technology-mediated assessment activities that provides a data-rich recording environment with which to investigate some of the assessment challenge components of case-based learning. The development of interactive cases is a task that requires formal implementation of assessment outcomes by teachers prior to their interaction with students. Unlike face-to-face interaction where teachers react to students' answers, the design of interactive case activities requires that teachers develop pre-planned answers and/or arguments to sustain a range of acceptable reasoning processes and answers.

The current work is situated in medical education where teachers are practising expert clinicians in their field who also have experience with the assessment of students' learning through cases. Expert clinical teachers can be assumed to have both a shared understanding of how to solve a case as well as a shared understanding of how to teach and assess these cases (Berliner 2001). This shared understanding of how to teach and develop students' learning is referred to as pedagogical content expertise (Berliner 1986). For this reason, we investigate how expert teachers conceptualize the notion of a good performance to a teaching case for a specific level of learner. Given that previous research has documented the variability in the problem-solving path between medical experts for the same problem (Elstein et al. 1978; Grant and Marsden 1988), we postulate that despite the fact that experienced clinical teachers may not solve a case exactly the same way, they share common expectations regarding the assessment of the case.

This paper begins by articulating the nature of the assessment challenge for case-based teaching approaches and proposing to address this challenge by studying the practical assessment knowledge of experienced teachers. Building on the idea that one important goal of case based teaching is to foster reasoning, we use the concept of key features (Page and Bordage 1995) to trigger teachers' shared understanding and expectations of a competent reasoning performance. We compare the expectations, answers, and reasoning processes for specific teaching cases by asking teachers to differentiate between key features according to whether they are critical, versus necessary or useful information for solving the case.

Case-based teaching practices focus on supporting the reasoning process

The use of cases is embedded in a number of instructional approaches—including case-based learning, case studies, and problem-based learning (PBL)—used across professions in higher education contexts (i.e. Medicine, Law, Business, Education). Even if these approaches differ in their ways of using and structuring teaching cases, they share some common key instructional goals and challenges (Moreno and Ortegano-Layne 2008; Savery 2006; Savin-Baden and Howell Major 2004; Sykes and Bird 1992; Williams 1992). More specifically, one shared key instructional goal of case-based learning approaches is to model and support the reasoning and decision-making processes involved in complex problem solving. Additionally, the assessment challenge we encountered is not solely due to the format and introduction of technologies into the equation, the literature on case-based teaching across professions has documented that this is a shared challenge even in face-to-face settings (Gijbels et al. 2005; Lundeberg and Yadav 2006; Savin-Baden 2004).

The concept of a "good answer" in these teaching cases is not straightforward: the same case may not be solved exactly the same way by the same person twice, nor is there only one way to reach a good answer; moreover, variability in the answers and ways to get to these answers—which is a desirable outcome when teaching with cases—adds to the challenge of measuring outcomes given that traditional assessment approaches treat variability in the answer or in the path as an error (Moss 1994; Schuwirth and der Vleuten 2006). A better understanding of the variability in the reasoning path or decision-making process supporting the inquiry process and the final answer(s) may be the key to inform the design of assessment guidelines and interactions for teaching cases.

Capturing pedagogical content knowledge of clinical teachers

Medical education represents a unique environment to study how teachers teach problem solving and clinical reasoning through the use of cases since the field has a long tradition of using cases to convey knowledge, and it has successfully implemented PBL—a type of case-based learning—for the past 30 years (Barrows 2000). Clinical teachers are experienced physicians who have expertise with both solving and teaching patient cases. These instructional experts are particularly adept at communicating their judgment strategies to an external audience. Unlike other types of experts, experienced teachers have the ability to break down the clinical reasoning process into comprehensible subunits and provide appropriate explanations for their decisions (Berliner 1986; Weiss and Shanteau 2003). This ability to provide contextual justifications of their reasoning and to understand how to explain it to an audience that is specific to their training level is conceptualized as pedagogical content knowledge (Gudmundsdottir 1991; Shulman 1986). We assume that this understanding of the content in relationship to the audience enables expert teachers to have assessment specific expectations for the reasoning process performance of each case.

Performance evidence in a computer-based context needs to be evaluated in terms of an absolute judgment or established benchmark instead of relative judgments that use students' comparison to rank the performance. This type of absolute judgment has shown to be influenced by the amount of experience of the person making the judgment (Govaerts et al. 2011). The development or setting of absolute criteria and standards prior to the task is referred to as a criterion reference-based approach to assessment (Worthen and Sanders

1987). Criteria and standards are two terms that are often used interchangeably but, technically speaking, a criterion refers to a property or characteristic of something by which its quality can be determined; whereas, a standard refers to the degree of quality recognized as adequate by an observer for some specific purpose (Sadler 2005). However, teachers' knowledge about assessment criteria for a specific task is often implicit; they know what a good performance looks like when they see it, but they cannot easily predict and articulate the criteria they use to make judgments (Sadler 1987).

Amongst the different methods used to set standards, exemplars are typically recommended for complex tasks (Shepard 1980). In the exemplar approach, a group of teachers use reviewed examples of student work to discuss and a reach consensus regarding criteria and standards for an open-ended task. The exemplar approach is a moderation approach to assessment that reveals the complexity of the judgments performed during assessment (Elmore 2002; Steele 1998).

Assessment of reasoning process: concept of key features

In medical education, a range of concepts and assessment tools exist to assess clinical reasoning (Nendaz and Tekian 1999). Tools like key features (Page et al. 1995) or script theory (Charlin et al. 2000), which are based on a schema theory framework (Patel and Arocha 1995; Rumelhart 1984), support the design of formal and standardized assessment tasks based on case scenarios and vignettes. Schema theory frameworks originating from the investigations into the nature of experts' knowledge representation and organization refer to how humans build scripts or cognitive templates through experience with recurring events, and has been influential in the development of assessment and computer-based learning systems (Marshall 1993; Schank and Abelson, 1977). The key feature approach to testing is based on the concept that most clinical encounters have a small number of important decisions that are instrumental in the successful resolution of the problem (Bordage and Page, 1987). Building on this robust and well-tested assessment concept we propose to use a different sampling procedure and development because our goal and context are for assessment for learning purposes.

The primary goal in the development of process measures for reasoning is to support learning, not to build standardized items for selection or certification purposes (Page and Bordage 1995). The type of cases for gatekeeping or selection purposes does not always correspond to the type of cases used in classrooms to develop specific competencies in learners. As a result, a statistical approach to the development of key features is impractical because it limits the selection of cases that have a well-agreed upon answer. Furthermore, this approach to the selection of important steps in the reasoning process loses its connection to the context and sequence of the context in which the decision-making process occurs. In standardized testing contexts, students are not given explanations or reasons to justify which answers they missed. In contrast, when cases are used in the classroom assessment context, they primarily aim at supporting the learning process and should provide transparent assessment criteria (Frederiksen and White 2004) that explain to learners the nature of the performance expected from them, a diagnostic of their performance, and how to address gaps in their performance (Sadler 1987). Assessment tasks in a competency-based approach are designed to trigger learning and provide tangible evidence for observers to judge and evaluate the work or performance.

Study

To investigate experienced teachers' assessment knowledge we study their judgments about assessment in a case-based teaching context. In this study we explore how experienced teachers (i.e., teachers who have years of experience with cases) conceptualize the notion of competent clinical reasoning performance for specific cases. To enable teachers to articulate explicitly their judgment about specific problem-solving tasks without drawing on their understanding of an unknown learner or misinterpretation of a verbal utterance, we ask them to first solve the case and then use a representation of their problem-solving performance as an exemplar to reflect on and to then identify criteria and standards for a specific audience. This approach enables us to situate the teachers' problem solving performance within the computer-based learning environment context. This approach also prevents the use of teachers' normative judgment and rating biases involved in using students' exemplars, and it builds on the premise that raters' judgments are grounded in a variety of their own experience and practice about a problem or situation (Gauthier and Czerski 2013).

Given previous research documenting variability in the problem-solving path between experts for the same problem (Elstein et al. 1978; Grant and Marsden 1988), we designed this study using the expert teachers' reflections about an exemplar based on their own performance to reveal their concepts of what constitutes a competent answer for a set of three introductory teaching cases. The teaching cases were selected as problems with a clear outcome or diagnosis, but where different processes and actions can be taken to reach that diagnosis. These types of problems provide an assessment challenge since arguments justifying the final diagnosis will vary, and the final diagnosis can be reached without proper inclusion and exclusion of other diagnoses and conditions.

Hypothesis

Hypothesis 1 When prompted to identify key features related to specific teaching cases, do teachers show better agreement regarding their expectations for competent reasoning performances despite their differences in how they actually solve these cases?

Hypothesis 2 When asked to categorize the key feature elements, do clinical teachers show better agreement on those that are critical, versus necessary or useful to know? We hypothesized that teachers' agreement on standards of a good reasoning performance would be higher for critical components, versus necessary or useful components related to the reasoning performance.

Method

A case study design (Yin 2009) was used to capture and provide in-depth analysis of the reasoning processes and assessment standards of five expert physician-teachers for a set of three endocrinology teaching cases. Each expert physician-teacher participated in two different tasks for each case. One task required solving a case while performing a think-aloud protocol, and the other task involved reflecting on a visual representation of their problem-solving performance of the case to identify important assessment criteria and standards for the case. This type of procedure is similar to previous studies on expert clinical reasoning, but instead of using an expert/novice framework to assess answers and study the types of mistakes made by novices (Elstein and Schwarz 2002), we focus on studying and identifying common "gold standards" or concepts of good answers for successful performances of typical teaching cases. The case study analysis aims to identify common concepts of competent problem solving performances and to set criteria and standards for specific teaching cases.

Participants

The five expert physician-teachers who participated in this study were all specialists in internal medicine and teach in the same Canadian medical school. The sample represents one-third of all the physician specialists teaching at this institution. For the purpose of this study, an expert was defined as someone with "prolonged or intense experience through practice and education in a particular field" (Ericsson 2006). We chose medical expert teachers who have an excellent grasp of the domain knowledge, sufficient amount of exposure to endocrinology cases, and are recognized by their peers as excellent teachers.

Problem-solving task

BioWorld (Lajoie 2009) is a computer-based learning environment that enables participants to interactively explore patient cases through the ordering of diagnostic tests, requesting information about vitals, visiting the library, or asking for a consult. Solving a case in this environment requires diagnosing the medical condition(s) affecting each patient. In addition to selecting a final diagnosis with a corresponding level of confidence, participants must add evidence that they collect during their case resolution, including patient symptoms, patient history, or diagnostic tests pertaining to their hypotheses. The structure of the environment is non-linear; participants can interact with the problem by selecting potential hypotheses, ordering test, checking vital signs and scrutinizing the patient problem in any sequence or order they want. We have tried to make the decision-making task as realistic as possible by providing a wide range of possible hypotheses (80), diagnostic tests (175) and textual elements (varies per case narrative) that can be selected as evidence by participants. When participants submit their final diagnosis, they are required to select and prioritize their evidence elements to support this final diagnosis.

Materials

Case material

A set of three endocrinology introductory level cases was used for this study. The cases represented typical instances of diabetes, hyperthyroidism, and pheochromocytoma. These three cases were selected because they represent prototypical cases taught in second year medical school. They were adapted from a textbook and tested by a content expert who was also an experienced teacher at the medical school. The cases were tested and validated by two other content experts. The cases were described as ill-defined problems since they have clear outcomes or diagnoses but no "best way" or sequence to reach the solution.

Questionnaires

Participants were administered a pre-questionnaire to document descriptive data related to participants' areas of expertise, recent clinical experience, and overall teaching experience. A post-questionnaire was administered to document their clinical and teaching experience with patient cases related to the ones they solved in this study.

Computer-based learning environment

We used the BioWorld (Lajoie 2009) computer-based learning environment to present participants with patient cases to solve. Participants can interact with the problem by selecting potential hypotheses, ordering tests, checking vitals, and scrutinizing the patient problem in any sequence or order they want. While solving the case, participants collect elements supporting their reasoning through an evidence palette, and they are asked to sort and prioritize the ones supporting their final diagnosis.

Procedures

Data collection occurred in two meetings. During the initial meeting, participants solved the three cases while performing a think-aloud protocol (Ericsson and Simon, 1993) using a computer-learning environment. Sound and video of the computer screen interaction were recorded using a screen capture program. In the first phase of analysis the computer log data and the think-aloud protocol were combined and analyzed to produce a sequential representation of the problem-solving performance for each case using a conversational analysis framework (Hutchby and Wooffitt 1998).

Participants were told to think out loud by verbalizing their thoughts as if they were giving a case presentation to an audience of second year medical students. The framing of the think-aloud protocol as a case presentation in this manner enables a conversational analysis perspective for these monologues where we focus on the intentions and meaning of the utterances and actions performed by participants (Gauthier 2012).

The visual representations of the decision space comprise summaries of thoughts in response to the case content in the computer-based learning environment and actions as they proceed to solve the case. As a result, the visual representations encompass elements that go beyond the content or interface presented to participants, including their thoughts about what else they would do when solving this case on the wards. These rich visual representations, like the section in Fig. 1, were used in this second meeting when participants were asked to first validate their visual representations for each case for accuracy and then to reflect on concrete examples of their thinking. Each line of the protocol could be accessed directly when doing a mouse-over or opening summaries. This reflection is similar to a retrospective think-aloud protocol where participants have the opportunity to add or comment on their previous performance (Ericsson and Simon, 1993). After validating the content, participants were asked to select and categorize which elements in their answer would correspond to critical, necessary, or useful elements to solve the case (see Fig. 2 below). A subsequent phase of analysis examined whether or not participants showed better agreement for critical elements, versus necessary or useful elements for all three cases.

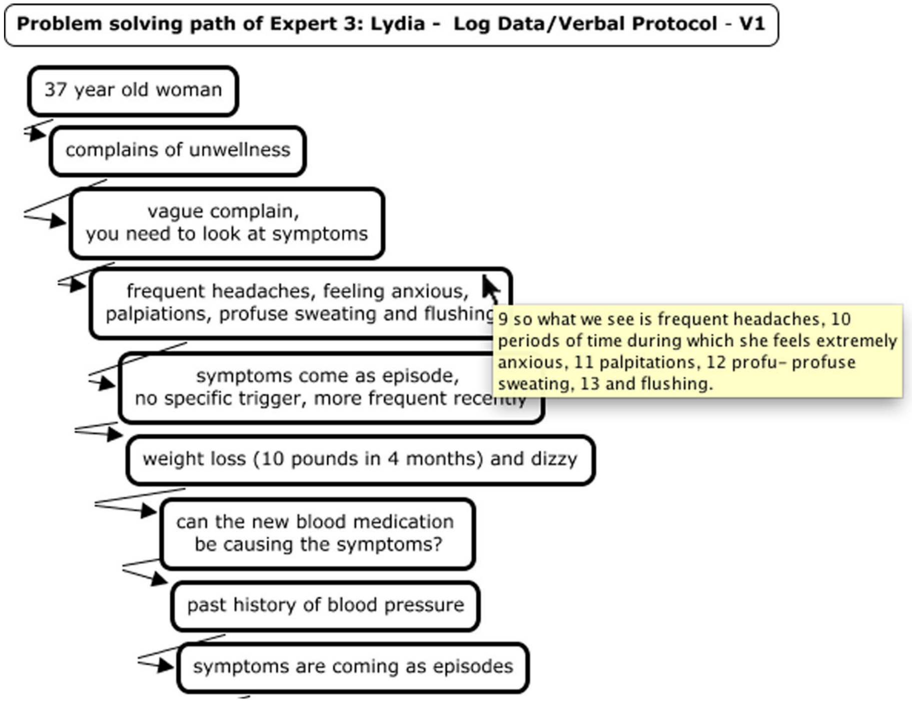

Fig. 1 Extract of the initial visual representation for Expert 3 for Case 1

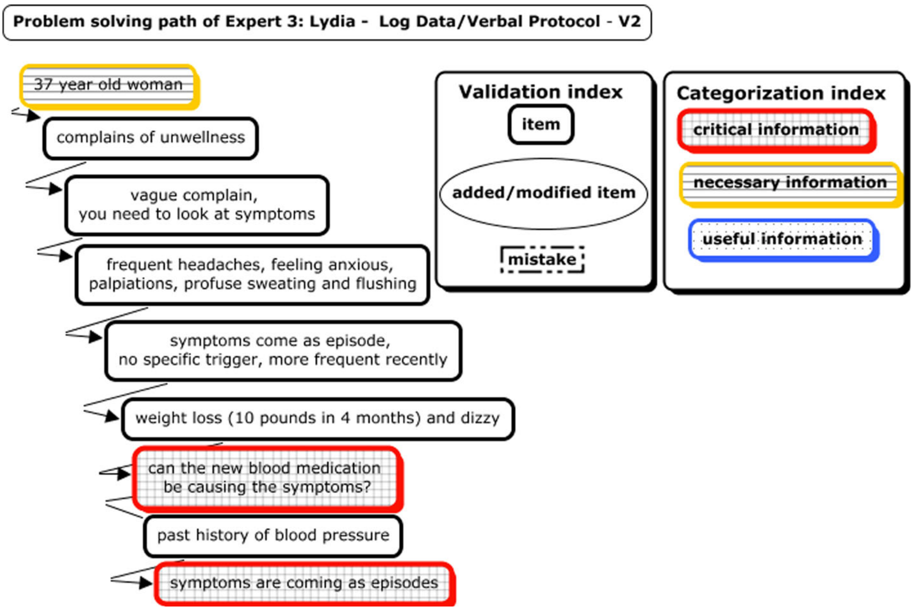

Fig. 2 Extract of the visual representation categorized by Expert 3 for Case 1

Data analysis

The analysis had two distinct phases. In the first phase of the analysis, prior to the second meeting with participants, the think-aloud and computer logs were analysed at the micro-level through a conversation analysis (CA) framework (Hutchby and Wooffitt 1998). The analysis of the protocol consisted of first segmenting and then creating episodes and short summaries of these episodes, which we refer to as elements of the problem solving process. The resulting sequential representations of the reasoning processes, anchored around units of meaning, were combined into a visual representation using Cmap tool (Novak and Cañas 2006). To segment the transcript we used individual computer log actions as well as the smallest complete ideas that were expressed verbally. The combination of lines into episodes corresponds to a set of ideas representing meaningful steps that contribute to the resolution of the case. For each case resolution we included all utterances and actions to insure transparency of the coding and accurate summarization of the problem solving performance. The reliability of the segmentation task was completed on all three cases with a Cohen's kappa of 80.94 %. For the development of episodes and their corresponding summaries that are elements in the visual representation, the comparison of coding from different researchers showed less than three different elements for the five cases coded in parallel (each case typically had between 40 and 65 elements).

After participants' identification and categorization of elements using these visual representations, a second phase of analysis focused on the convergence of elements identified by clinical teachers as being critical, necessary or useful for each case. Each identified element was compared in terms of sequence and content to the four other representations (i.e., in Case 1, all participants talk about and select weight loss of the patient). When a minimum of three experts had identified similar elements as having significance (necessary, necessary or useful) the elements were combined into key elements. For example, in the teaching case on pheochromocytoma below, the age of the patient was categorized as an important element by all five experts as shown by the patterns of elements in Fig. 3a; therefore, these elements were combined into a common key element as seen in Fig. 3b below. When combining these common key elements for the five experts during the last phase of analysis we obtained a contextual sequence of common key elements and individual variability for the reasoning process of each case.

Results

Overview

Participants solved the three teaching cases differently even though they submitted similar final diagnoses for each case; they ordered different tests, identified different elements as being significant, and explored different potential hypotheses. To understand participants' assessment concepts of good reasoning and problem-solving ability, we asked them to make differential selections of the key elements supporting their final answers and the ones supporting their reasoning processes. As mentioned in our first hypothesis, participants show a better consensus on the key elements they select to support good reasoning processes than the ones they select to support a good final answer for each case, as shown in Table 1. However, our second hypothesis needs to be rejected because they do not show any consensus regarding which key element is critical, versus necessary or useful.

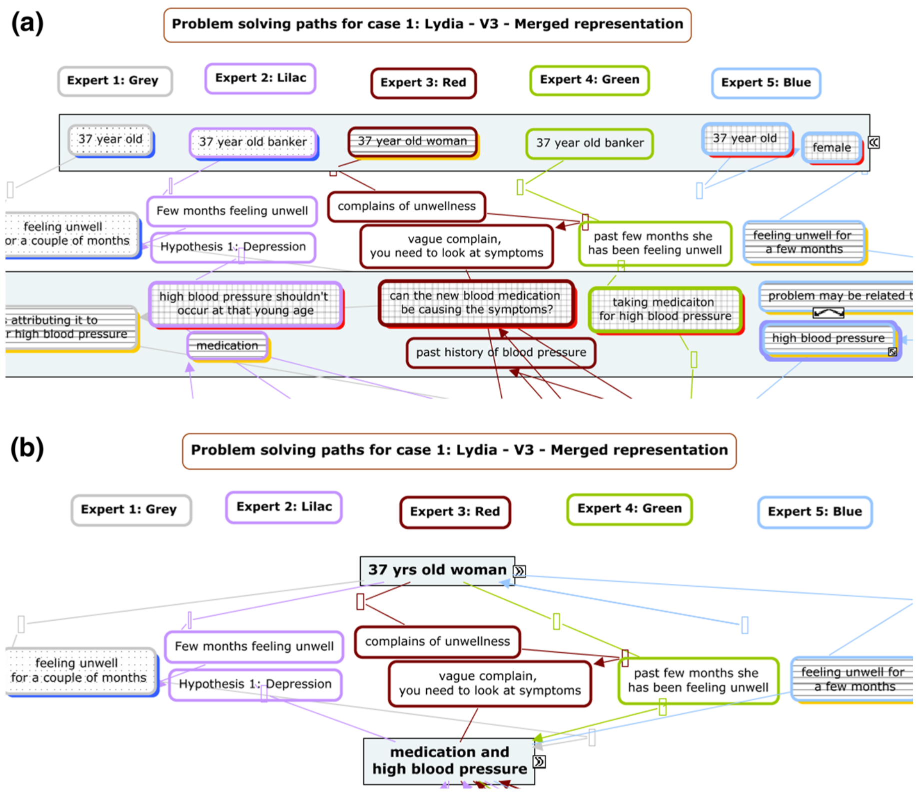

Fig. 3 Extract of the merged protocols from 5 experts for Case 1

Nevertheless, if we ignore the differential categorization, participants do agree on elements that have some importance versus the ones that are not relevant Table 2.

The overview of consensus rate focused on answers or the processes related to each case. Table 1 shows the consensus rate for answers varying, with 42 % for Case 1, 55 % for Case 2, and 78 % for Case 3. When comparing the elements supporting the assessment of the reasoning process, we found an agreement level of 79 % for Case 1, 88 % for Case 2, and 87 % for Case 3. The agreement or consensus was defined as more than 50 % of participants listing or selecting elements as being key to support successful reasoning for the specific case .

Multiple solution paths and key elements

Participants solved three teaching cases using different paths and chose different key elements to support and justify their final answers. No common pattern of interaction could be found when comparing participants' interactions with the cases as captured by the

Table 1 Consensus rate of elements supporting the answer versus reasoning process

Consensus rate	Case 1 (%)	Case 2 (%)	Case 3 (%)
Answer elements	42	55	78
Process elements	79	88	87

computer logs or the sequential think-aloud protocol. Expert teachers took different paths to solve each teaching case while submitting similar answers for each case (except for one participant on Case 1). For example, as documented in the computer logs, participants ordered between 12 and 20 different diagnostic tests for which they only had two in common. Participants did not select identical elements to support their final answer, nor did they take the same amount of time to solve the cases in this study. As for the exception in case one, Expert teacher 2 admittedly listed the correct diagnosis of Pheochromocytoma as his initial hypothesis; but due to the very low probability of this disease, he argued that the most likely diagnosis to be made by a physician or a student in this context should be panic attacks.

The key elements selected by participants to support their final diagnosis varied. Each of the expert participants selected and ranked relatively different key elements supporting their answer for Case 1. The number of key elements and the nature of these elements varied. The only common key element that was selected was anxiety combined with palpitation, sweating, and flushing. Two of the five experts, Experts 1 and 2, put more emphasis on diagnostic tests as being key elements supporting the answer, while the two others considered them as secondary or not really an important part of the solution despite having ordered them during their interaction with the case. This variation in the ranking, number, and nature of the key elements creates a problematic assessment blueprint to score or determine acceptable elements supporting the reasoning for this teaching case.

Common key elements selected to support the reasoning process

The analysis of the assessment evidence selected as critical, necessary, or useful revealed that participants showed agreement on what they selected as being meaningful evidences versus the evidences that they did not select at all. Overall, participants selected <50 % of the total number of evidences presented in the visual representation. Table 2 shows the number of evidences they categorized in proportion to ones that represented shared evidences for each case.

To illustrate the nature and sequence of the common elements of the solving process for each case we provide an example of the common elements selected for the pheochromocytoma case. Only 30 of the 114 categorized elements were not part of the common elements, and a third of these 30 categorized elements are associated with Expert 5, who did not submit the same diagnosis—an exception worth noting. Given that Expert 5 had a different final diagnosis of panic attack, we could have expected a greater difference in the selection of key evidences for the reasoning process, but this was not the case as his reasoning had a lot in common with the four other experts.

At the beginning of the case, all experts mentioned the age of the patient while four of them categorized her age as a key element. Inside the protocol of these elements, Experts 2 and 5 explicitly commented on the impact of age for the evaluation of the current problem. Then, even if all of the experts mentioned the complaint of feeling unwell, it does not correspond to a key element since only two experts categorized this symptom as a key

Table 2 Total number of categorized elements compared to shared categorized elements

	Case 1	Case 2	Case 3
Categorized elements	144	121	143
Shared categorized elements	114	107	125
Consensus rate (%)	79	88	87

element. The next common key element for this case related to the patient's high blood pressure and the medication she was taking. All participants mentioned these two elements, either as one or two separate elements and categorized them. The next common key element identified was the association of the new blood pressure medication as a potential explanation for the symptoms, followed by symptoms experienced in episodes; the linking of the different symptoms of palpitations with profuse sweating and flushing together; and finally, the 10-pound weight loss in four months. All participants mentioned or selected the main hypothesis of hyperthyroidism and listed pheochromocytoma as an alternative. Other hypotheses were listed by three of the participants but they did not match, nor were they identified as being key elements by more than three of the participants. When reviewing the chart, all but one participant talked about and categorized the elevated blood pressure and pulse rate as common elements. All of them expressed the need to test for baselines and ordered blood glucose and electrolyte diagnostic tests.

Overall, when comparing the five individual problem-solving representations for each case, we found that common categorized elements could be summarized by 11–14 common elements for each case. These common elements explained, on average, 85 % of the total categorized elements identified as important in the problem-solving process. In other words, even if participants did not follow the same path or steps to solve the patient cases, they had a common understanding of key elements when reflecting on their problem-solving processes for each case.

No agreement of the differential importance of key elements

Contrary to our second hypothesis, despite agreeing on which elements represented key elements, participants did not show a higher level of agreement for elements categorized as critical to know, versus the ones categorized as necessary to know or useful to know. We tried to understand which element was considered critical, and to examine why there was not one element that more than two experts agreed upon. There was no convergence on any critical element selected by the five experts for Case 1. Contrary to our hypothesis, the differential categorization of elements emphasized the variability that occurred among expert teachers rather than reduced it. They did not select similar elements nor did they select a comparable amount of elements as being critical. Additionally, there was no consensus on the categorization of elements for other categories or cases.

Conclusion and discussion

To explore new ways to support the assessment of teaching cases aligned with the instructional goals and purposes of case based teaching, we examined how experienced clinical teachers conceptualize the notion of good reasoning performances for specific cases. Unlike other expertise studies of clinical reasoning, we did not collect experts' performances to define or discriminate for optimal reasoning strategies or characteristics, but we used these performances as a basis of expert teachers' reflection for case-based assessment.

For this case study, an in-depth protocol analysis of problem-solving performances was conducted as well as an analysis of teachers' reflections. Using the concept of key features (Page and Bordage 1995), we collected and analyzed contextual expert teachers' contextual reflections on their assessment expectation and judgment of good reasoning instead of their own concept of the best answer or performance for each case. We compared what teachers identified as being the key elements to support their inferences about the judgment of successful reasoning for each case. More specifically, we examined how five participants independently solved three cases before reflecting on the key elements supporting the assessment of answers and the ones supporting reasoning processes for these teaching cases. Results suggest that despite solving cases differently expert teachers show higher consensus on assessment criteria supporting the reasoning process than on the ones supporting the final answers. However, results also point out that expert teachers do not share the same standards regarding assessment evidence; they do not agree on the relative importance of any of the key elements selected.

Our findings indicate that incorporating some types of process measures in case based learning may be relevant, as well as proposing ways to acknowledge variability of decision making in context. These findings further suggest that the development of case-specific assessment models may enrich the validity and transparency of the assessment process for both teachers and learners. In terms of implications for the development of computer-based learning environments, our findings suggest the pertinence of incorporating more than one pedagogical expert, not only one content expert, including the potential of using teachers' reflections to develop and update these case-specific models.

The level of agreement between clinical teachers on key elements suggests that it may be feasible and relevant to develop assessment models for clinical reasoning that incorporate reasoning processes along with other potential outcome measures. Given that one of the key instructional goals of the case based teaching approach is to foster and model the reasoning process, it is important to develop methods to assess and support the development of this process. In previous work we have documented intra- and inter-expert variability when asking experts to justify their answer (Gauthier et al. 2012), but this variability may have been due to a memory issue or the backward type of reasoning involved when justifing answers.

In terms of method, the use of visual representations has the potential to inform the contextual nature of reasoning that occurs as participants collect data, generate hypotheses, and test hypotheses before deciding on a final hypothesis. The representation for each case provides a comprehensive repertoire of shared and divergent decisions throughout the problem-solving task for each case. As emphasized by Voss and Post (1988), the display of argumentation could be a way to judge the quality of solutions for problems that have no clear answers since there are no universal criteria or absolute truths for many of these problems occurring in different contexts.

Anchoring assessment design in empirical sampling of more than one expert's performance might prevent construct-irrelevant variance, i.e., construct variance related to the variance in the data that is not relevant to the interpretation of what we are trying to assess (Messick 1989). Showing the variance that exists at the level of "competent" practitioners can challenge the notion of what is and is not appropriate to evaluate. For example, when assessing learners' performance on their reasoning process, grading of elements corresponding to common key elements will be different than on aspects that correspond to irrelevant elements performed by expert teachers but not selected as relevant. In other words, when learners repeat similar sub-optimal reasoning that experts might also have exhibited, the grading of the performance might lead to more nuanced assessment or feedback. Overall, potential use of the case-specific models could improve what is referred to as systemic validity (Frederiksen and Collins 1989) as it has the potential to inform both the learners and the teacher about the nature of competent performance. If learners can understand where their relative weaknesses are, it improves the transparency of the assessment process and enables a better evaluation of the validity claim and the corresponding inference of proficiency related to its scoring in small-scale educational settings (Kane 1992). Understanding the meaning of the data in relationship to both the global and contextual nature of the performance standards used is necessary for formulating clear and transparent arguments.

Despite the richness of the contextual data present in the proposed assessment models, the lack of common prioritization or categorization of key elements may indicate the contextual nature of inferences made by teachers when assessing specific groups of students. Some research on teachers' judgment suggests that their inference about students' performance is always influenced by the cohort effect (Wiliam 1996). Another explanation of this lack of agreement on the relative importance of each individual element could be due to the dichotomous nature of the judgment of standards that teachers usually perform (Silber et al. 2004).

As the development of computer-based learning environments is moving forward, this study suggests that the use of cognitive task analysis by content experts may not be sufficient to produce relevant content and assessment. As suggested by the adaptive expertise model (Ericsson et al. 1993; Hatano and Inagaki 1986; Mylopoulos and Regehr 2007), experts' reflections on their performances may be more useful than simply focusing on their performance per se to build explicative models of clinical reasoning. Lastly, the design of content and structures to support that content may need to include teachers at a different level to develop and revise assessment models that are specific and contextual.

Limitations

The study findings should be interpreted with caution in appreciation of a number of limitations. The number of participants is small due to the challenge of recruiting expert teachers with the necessary physician qualifications. Similarly, the small number of cases presented in this study is related to the scarcity of medical expertise to develop the cases because each case requires a significant investment of time and resources. The cases were not designed to be representative of the field of internal medicine, nor do they significantly cover the topic of endocrinology that limits the implication of the findings.

References

Albanese, M. A., & Mitchell, S. (1993). Problem-based learning: A review of literature on its outcomes and implementation issues. *Academic Medicine, 68*(1), 52–81.

Barrows, H. S. (2000). *Problem-based learning applied to medical education*. Springfield, IL: Southern Illinois University School of Medicine.

Barrows, H. S., & Tamblyn, R. M. (1980). *Problem-based learning: An approach to medical education*. New York: Springer Publishing Company.

Berliner, D. (1986). In pursuit of the expert pedagogue. *Educational Researcher, 15*(7), 5–13.

Berliner, D. C. (2001). Learning about and learning from expert teachers. *International Journal of Educational Research, 35*(5), 463–482.

Bordage, G., & Page, G. (1987). An alternate approach to PMPs, the key feature concept. In I. Hart & R. Harden (Eds.), *Further developments in assessing clinical competence* (pp. 57–75). Montreal, QC: Can-Heal Publications.

Brookhart, S. M. (2003). Developing measurement theory for classroom assessment purposes and uses. *Educational Measurement: Issues and Practice, 22*(4), 5–12.

Charlin, B., Tardif, J., & Boshuizen, H. P. A. (2000). Scripts and medical diagnostic knowledge: Theory and applications for clinical reasoning instruction and research. *Academic Medicine, 75*(2), 182–190.

Cook, D. A. (2007). Web-based learning: Pros, cons and controversies. *Clinical Medicine, Journal of the Royal College of Physicians, 7*, 37–42.

Dochy, F., Segers, M., Van den Bossche, P., & Gijbels, D. (2003). Effects of problem-based learning: A meta-analysis. *Learning and Instruction, 13*(5), 533–568.

Elmore, R. (2002). *Bridging the gap between standards and achievement: Report on the imperative for professional development in education*. Washington, DC: Albert Shanker Institute.

Elstein, A. S., & Schwarz, A. (2002). Clinical problem solving and diagnostic decision making: Selective review of the cognitive literature. *British Medical Journal, 324*(7339), 729–732.

Elstein, A., Shulman, L., & Sprafka, S. (1978). *Medical problem solving*. Cambridge, MA: Harvard University Press.

Ericsson, K. A. (2006). An introduction to the Cambridge handbook of expertise and expert performance: Its development, organization, and content. In K. A. Ericsson, N. Charness, P. J. Feltovich, & R. R. Hoffman (Eds.), *The Cambridge handbook of expertise and expert performance* (pp. 3–20). Cambridge, UK: Cambridge University Press.

Ericsson, K. A., Krampe, R. T., & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review, 100*(3), 363–406.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data*. Cambridge, MA: MIT Press.

Frederiksen, J. R., & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher, 18*(9), 27–32.

Frederiksen, J., & White, B. Y. (2004). Designing assessments for instruction and accountability: An application of validity theory to assessing scientific inquiry. *Yearbook of the National Society for the Study of Education, 103*(2), 74–104.

Gauthier, G. Conway, J. M. and Taylor, R. (2012). *Variability in the expert solution for case based learning scenarios: Reliability issues*. Paper presented at the Canadian Society for the Study of Education (CSSE). Waterloo, Ontario.

Gauthier, G. (2012). *Teach aloud: A modified version of the think-aloud protocol to study the teaching of clinical reasoning*. Paper presented at the Qualitative Health Research (QHR). Montreal, Quebec.

Gauthier, G. and Czerski, M. (2013). *Visual representation of teaching discourse: Tool enabling "reflection-on-action"*. Paper presented at the 12th annual advances in qualitative methods (AQM) conference, Edmonton, Alberta.

Gijbels, D., Dochy, F., & Bossche, P. V. (2005). Effects of problem-based learning: A meta-analysis from the angle of assessment. *Review of Educational Research, 75*, 27–61.

Govaerts, M. J. B., Schuwirth, L. W. T., van der Vleuten, C. P. M., & Muijtjens, A. M. M. (2011). Workplace-based assessment: Effects of rater expertise. *Advances in Health Sciences Education, 16*(2), 151–165.

Grant, J., & Marsden, P. (1988). Primary knowledge, medical education and consultant expertise. *Medical Education, 22*, 173–179.

Greenhalgh, T., & Hurwitz, B. (1999). Narrative based medicine: Why study narrative? *BMJ, 318*(7175), 48–50.

Gudmundsdottir, S. (1991). Ways of seeing are ways of knowing. The pedagogical content knowledge of an expert English teacher. *Journal of Curriculum Studies, 23*(5), 409–421.

Hatano, G., & Inagaki, K. (1986). Two courses of expertise. In H. A. H. Stevenson, & K. Hakuta (Eds.), *Child development and education in Japan* (pp. 262–272). New York: W. H. Reeman and Company.

Hmelo, C. (1998). Problem-based learning: Effects on the early acquisition of cognitive skill in medicine. *The Journal of the Learning Sciences, 7*(2), 173–208.

Hmelo-Silver, C. E., Maher, C. A., Agne, G., Palius, M., & Derry, S. J. (2010). The video mosaic: design and preliminary research. *Proceeding of ICLS* (pp. 425-426).

Hutchby, I., & Wooffitt, R. (1998). *Conversation analysis: Principles, practices and applications*. Oxford, UK: Blackwell Publishers Inc.

Kane, M. T. (1992). An argument-based approach to validation. *Psychological Bulletin, 112*, 527–535.

Lajoie, P. S. (2009). Developing professional expertise with a cognitive apprenticeship model: Examples from avionics and medicine. In K. A. Ericsson (Ed.), *The Development of professional performance: Approaches to objective measurement and designed learning environments* (pp. 61–83). Cambridge, UK: Cambridge University Press.

Lundeberg, M. A., & Yadav, A. (2006). Assessment of case study teaching: Where do we go from here? Part I. *Journal of College Science Teaching, 35*(5), 10–13.

Marshall, S. (1993). Assessing schema knowledge. In N. Frederiksen, R. Mislevy, & I. Bejar (Eds.), *Test Theory for a new generation of tests* (pp. 155–180). Hillsdale, NJ: Erlbaum.

Merseth, K. K. (1994). *Cases, case methods, and the professional development of educators*. Washington, DC: ERIC Clearinghouse on Teaching and Teacher Education (BBB30990).

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18*(2), 5–11.

Moreno, R., & Ortegano-Layne, L. (2008). Using cases as thinking tools in teacher education: The role of presentation format. *Educational Technology Research and Development, 56*, 449–465.
Moss, P. A. (1994). Can there be validity without reliability? *Educational Researcher, 23*(2), 5–12.
Mylopoulos, M., & Regehr, G. (2007). Cognitive metaphors of expertise and knowledge: Prospects and limitations for medical education. *Medical Education, 41*(12), 1159–1165.
Nendaz, M. R., & Tekian, A. (1999). Assessment in problem-based learning medical schools: A literature review. *Teaching and Learning in Medicine, 11*(4), 232–243.
Novak, J. D., & Cañas, A. J. (2006). *The Theory Underlying Concept Maps and How to Construct Them. Technical Report IHMC CmapTools 2006-01*. Florida Institute for Human and Machine Cognition.
Page, G., & Bordage, G. (1995). The medical council of canada's key features project: A more valid written examination of clinical decision-making skills. *Academic Medicine, 70*(2), 104–110.
Page, G., Bordage, G., & Allen, T. (1995). Developing key-feature problems and examinations to assess clinical decision-making skills. *Academic Medicine, 70*(3), 194–201.
Patel, V. L., & Arocha, J. F. (1995). Cognitive models of clinical reasoning and conceptual representation. *Methods of Information in Medicine, 34*(1–2), 47–56.
Rumelhart, D. E. (1984). Schemata and the cognitive system. In R. S. Wyer & T. K. Srull (Eds.), *Handbook of social cognition* (Vol. 1, pp. 161–188). Hillsdale, NJ: Erlbaum.
Sadler, D. R. (1987). Specifying and promulgating achievement standards. *Oxford Review of Education, 13*(2), 191–209.
Sadler, D. R. (2005). Interpretations of criteria-based assessment and grading in higher education. *Assessment and Evaluation in Higher Education, 30*(2), 175–194.
Savery, J. R. (2006). Overview of problem-based learning. *The Interdisciplinary Journal of Problem-based Learning, 1*(1), 9–20.
Savin-Baden, M. (2004). Understanding the impact of assessment on students in problem-based learning. *Innovations in Education and Teaching International, 41*(2), 221–233.
Savin-Baden, M., & Howell Major, C. (2004). *Foundations of problem-based learning*. Maidenhead: SRHE/Open University Press.
Schank, R. C., & Abelson, R. P. (1977). *Scripts, plans, goals and understanding: An inquiry into human knowledge structures*. Hillsdale, NJ: Erlbaum.
Schön, D. A. (1987). *Educating the reflective practitioner*. San Fransico: Jossey-Bass.
Schuwirth, L. W., & van der Vleuten, C. P. (2006). A plea for new psychometric models in educational assessment. *Medical Education, 40*(4), 296–300.
Shepard, L. (1980). Standard setting: Issues and methods. *Applied Psychological Measurement, 4*(4), 447–467.
Shepard, L. A. (2006). Classroom assessment. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 623–646). Washington, DC: National council on measurement in education and American council on education/praeger.
Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher, 15*(2), 4–14.
Shulman, L. (1992). Toward a pedagogy of cases. In J. H. Shulman (Ed.), *Case methods in teacher education* (pp. 1–30). New York: Teachers College Press.
Silber, C. G., Nasca, T. J., Paskin, D. L., Eiger, G., Robeson, M., & Veloski, J. J. (2004). Do global rating forms enable program directors to assess the ACGME competencies? *Academic Medicine, 79*(6), 549–556.
Steele, B. (1998). Exemplars: An interpretive and collaborative framework for assessment that works. *English Quarterly, 30*(1–2), 82–86.
Sykes, G., & Bird, T. (1992). Teacher education and the case idea. In G. Grant (Ed.), *Review of research in education* (Vol. 18, pp. 457–521). Washington, DC: American Educational Research Association.
Voss, J. F., & Post, T. A. (1988). On the solving of ill-structured problems. In M. T. H. Chi, R. Glaser, & M. J. Farr (Eds.), *The nature of expertise* (pp. 261–285). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.
Weiss, D. J., & Shanteau, J. (2003). Empirical assessment of expertise. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 45*(1), 104–116.
Wiliam, D. (1996). Meanings and consequences in standard setting. *Assessment in Education: Principles, Policy and Practice, 3*(3), 287.
Williams, S. M. (1992). Putting case-based instruction into context: Examples from legal and medical education. *The Journal of the Learning Sciences, 2*(4), 367–427.
Worthen, B. R., & Sanders, J. R. (1987). *Educational evaluation: Alternative approaches and practical guidelines*. New York: Longman.
Yin, R. (2009). *Case study research: Design and methods* (4th ed.). Thousand Oaks, California: Sage.