Available online at www.sciencedirect.com

ScienceDirect

Health Professions Education 5 (2019) 237–247

www.elsevier.com/locate/hpe

Guidelines for Creating Written Clinical Reasoning Exams: Insight from a Delphi Study

Évelyne Cambron-Goulet[a,*], Jean-Pierre Dumas[b], Édith Bergeron[c], Linda Bergeron[c], Christina St-Onge[c]

[a]*Département des sciences de la santé communautaire, Université de Sherbrooke, 3001 12e Avenue Nord, Sherbrooke, Québec, Canada J1H 5N4*
[b]*École de réadaptation, Université de Sherbrooke, 3001 12e Avenue Nord, Sherbrooke, Québec, Canada J1H 5N4*
[c]*Chaire de recherche en pédagogie médicale Paul Grand-Maison de la Société des médecins de l'Université de Sherbrooke, 3001 12e Avenue Nord, Sherbrooke, Québec, Canada J1H 5N4*

Received 1 June 2018; received in revised form 13 August 2018; accepted 2 September 2018
Available online 6 September 2018

Abstract

Context: Clinical reasoning is an essential skill to be learned by medical students, and thus requires to be assessed. Although written exams are widely used as one of the tools to assess clinical reasoning, there are no specific guidelines to help an exam writer to develop good clinical reasoning assessment questions. Therefore, we conducted a modified Delphi study to identify guidelines for writing questions that assess clinical reasoning.
Methods: Participants were identified from: 1) the literature on clinical reasoning (i.e., people who wrote about clinical reasoning and assessment), 2) the people responsible for assessment in Canadian medical faculties, and 3) a snowball sampling strategy. Thirty-two question-writing guidelines were drawn from the literature and adapted by the team members. Participants were asked to indicate on a ten-point Likert scale their perceived importance of each guideline, and, starting in the second round, the relevance of each guideline in five assessment contexts. A total of three rounds were conducted.
Results: Response rates were 24%, 57%, and 62% for each round, respectively. Consensus about the importance of the guidelines (interquartile range < 2.5) was reached for all but four guidelines. Four guidelines were identified as important (median ≥ 9 on ten-point scale): the question should be based on a clinical case, the question represents a challenge achievable for the student, the correction scale (i.e., scoring grid) is explicit, and a panel of experts revises the questions.
Conclusion: A large number of guidelines seem relevant for written-exam clinical reasoning assessment questions. We are considering grouping those guidelines into categories to create a simple tool for use by medical educators in the design of written-exam clinical reasoning assessment questions. The next step will then be to collect evidence of validity about this tool: Does it really help to build questions that assess clinical reasoning?

Keywords: Clinical reasoning; Assessment; Written-exam questions; Guidelines; UGME

*Correspondence to: Centre intégré de santé de services sociaux de la Montérégie-Centre, 1255 rue Beauregard, Longueuil, Québec, Canada, J4K 2M3 Canada.
E-mail addresses: evelyne.cambron-goulet@usherbrooke.ca (É. Cambron-Goulet), jean-pierre.dumas@usherbrooke.ca (J.-P. Dumas), edith.bergeron2@usherbrooke.ca (É. Bergeron), linda.bergeron@usherbrooke.ca (L. Bergeron), Christina.st-onge@usherbrooke.ca (C. St-Onge).
Peer review under responsibility of AMEEMR: the Association for Medical Education in the Eastern Mediterranean Region

https://doi.org/10.1016/j.hpe.2018.09.001

1. Introduction

Clinical reasoning is at the heart of physicians' daily practice of medicine.[1] Several authors have mapped its manifestation in novices versus experts (e.g., Schmidt & Boshuizen),[2] or debated the underlying mechanisms that make it possible for physicians to get the "correct" diagnosis.[3,4] Notwithstanding the many debates on what exactly is clinical reasoning[5] and how we can foster it in trainees, most clinicians would agree that clinical reasoning encompasses the thoughts and decision-making processes that enable physicians to choose the most relevant actions in a given context.[6] Medical-training program directors are increasingly recognizing the importance of teaching clinical reasoning and, as such, are embedding the teaching of clinical reasoning within their curricula.[7–9] Since assessment drives learning, and given the importance of clinical reasoning in future practice, administrators and directors from both undergraduate and postgraduate medical training programs are struggling to identify, develop and administrer novel strategies that will allow them to assess trainees' clinical reasoning.[10]

Traditionally, clinical reasoning has been assessed in the context of clinical rotations or in simulated clinical context, such as the Objective Structured Clinical Examinations (OSCEs).[11] However, these strategies are resource-consuming, and are often introduced later in the curriculum. In order to introduce less resource-consuming assessment strategies earlier in the curriculum to assess trainees' clinical reasoning, and therefore identify students who would need remediation before entering the clinical setting,[12] some administrators and program directors are considering the use of written examinations.

Different written-exam formats have been used to assess clinical reasoning, for example, key-feature questions, extended-matching questions, multiple-choice questions (MCQ),[13–15] as well as script concordance tests (SCT) and modified essay questions.[16,17] One format does not appear to be better than another at assessing clinical reasoning. Indeed, evidence suggests that the quality of the items is more important than their format, that is, how a question is written is more important than it being an MCQ, a Short-answer Question (SAQ) or an SCT.[14,18] This is why having guidelines for writing items to assess clinical reasoning is essential.

There are some general guidelines for written examination questions (i.e., guidelines such as "never using the negative mode to write a question"),[19] and some guidelines for specific types of questions used to assess clinical reasoning (such as script concordance tests).[28] However, there is a lack of specific guidelines for developing clinical reasoning assessment questions for learners. Considering the lack of guidelines for the development of written-exam clinical reasoning assessment questions, this study is a first step in the development and validation of a tool that would help question writers to develop written-exam clinical reasoning questions. The purpose of this study is thus to identify, with the help of a group of experts, guidelines used to develop written-exam clinical reasoning assessment questions in different contexts, such as level of training and purpose of the assessment.

2. Method

2.1. Design

We conducted a Delphi study to identify, via a consensus of experts, the guidelines that should be used to guide the development written-exam clinical reasoning assessment questions. The purpose of a Delphi method is to achieve a consensus with a panel of experts on a particular topic.[20] It is conducted in several consecutive rounds during which the participants can add new elements to the group reflection and then adjust their positions according to the other participants' responses.[21] This method has the advantage of fostering a consensus by giving all participants access to the panel's responses in each round.[20] Moreover, the method provides a means for consulting a large number of experts without having to gather them in the same geographic location. It also avoids direct confrontation of the experts[20] and makes it possible to question experts on various aspects of a single theme during one consultation, as well as allows them time for reflection between each round.[21]

More specifically, the objective of this study was to identify guidelines for writing items to assess clinical reasoning at different levels of training (preclinical, clerkship, and residency) and according to different assessment purposes (summative vs. formative). This project received the approval of our local IRB (2014-53).

2.2. Participants and recruitment

2.2.1. Identifying experts

The process to identify potential participants was carried out in four steps. At the first step, three team members (ECG, LB and CSTO) consulted their respective networks to identify individuals known to have expertise in clinical reasoning.

At the second step, one team member (LB) conducted a literature review to identify health professions education researchers in the field of clinical reasoning (subsequently refered to as researchers).

Inclusion criteria included being a currently active researcher (publication within the last year) and having published on the topic of clinical reasoning. The list of presenters at the Second Montréal International Conference on Clinical Reasoning in 2014 was also consulted. LB drafted a list of potential participants for revision by two team members (ECG and CSTO).

At the third step, e-mails were sent to the vice-deans of undergraduate and postgraduate medical education at Canada's 17 medical schools, asking them to provide a list of individuals in charge of assessment at their institutions (subsequently refered to as users). Eight of the medical schools responded.

Lastly, at the fourth step, we adopted a "snowball strategy" by asking round-one participants to identify experts whom they considered to be influential actors in the field of clinical reasoning and its assessment. The experts who were referred by participants at least two weeks before the end of round one and who met our original criteria were integrated into the data collection for this round; otherwise, they were added to round two.

2.3. Materials and methods

We carried out a Delphi study design, beginning round one with a pre-determined list of guidelines to be rated according to their importance for assessing clinical reasoning. A total of three rounds were conducted. The objective of round one was to complete the list of guidelines with new guidelines that the experts would identify as being missing from our list. For rounds two and three (as described below), the objectives were to obtain a consensus on the relevance of each guideline and to clarify the context (level of training and purpose of the assessment) in which these item-writing guidelines would be important.

2.3.1. Identification of guidelines

We began this process with a list of four references (Frey et al., 2005; Haladyna et al., 1989; Haladyna et al., 2002; Auger et al., 2000) reporting general item-writing guidelines. From these references, we listed 107 guidelines. Following the method used by Vachon Lachiver, St-Onge, Cloutier, and Farand (2018)[22], team members ECG and LB reviewed the literature via Google Scholar to search for item-writing guidelines specific to written-exam clinical reasoning assessment questions. They did not find any such guidelines, but found guidelines for specific question formats (e.g., SCT) that claim to assess clinical reasoning.[19,23–29] The initial list of 107 guidelines was completed with the guidelines for these specific question formats (n = 101), for a total of 208 guidelines.

We then grouped together items that were related to each other (e.g., guidelines about how to build a question, guidelines about content of the questions, guidelines about the level of students, etc.). This allowed us to eliminate guidelines that were redundant (n = 87). Then, through an iterative process, and by having CSTO tie-break any discrepancies in their judgements, they (ECG and LB) identified guidelines that could be applied in the context of the written assessment of clinical reasoning (e.g., guidelines that concerns *true or false items* were excluded, guidelines that concerns the format of the exams were excluded, etc.; n = 59 guidelines excluded). The list was narrowed down to 62 guidelines during this process.

The remaining 62 guidelines were reviewed to ensure they were 1) mutually exclusive, 2) operational, 3) had a single-focus (i.e., the guidelines had to focus only on one idea, otherwise it was divided, see guidelines 18 and 19 in Appendix B for an example), and 4) were free of ambiguities. During this last step of excluding and/or splitting the guidelines we ended up with 28 guidelines.

We decided to group the guidelines according to their topic. We then pilot tested the questionnaire with two volunteers having experience in the field of assessment or clinical reasoning (who would not participate in the study). The suggestions they made to improve the reading of certain guidelines and to add certain other guidelines were discussed between CSTO, ECG and LB. A total of 32 guidelines, grouped into four sections (see Appendix B), were included in the initial questionnaire. All documents were created in both French and English. The questionnaires for each round are described below. Note that during the translation process of the questionnaire, the term, *correction* was used in the questionnaires instead of the more familiar *scoring*. This would have changed: *Correction* should have been *Scoring*; *correction key* should have been *marking key*, *correction process* should have been *process of scoring and grading*.

2.4. General procedure

A letter stating the study objective and process was sent out by email to all potential participants. Participants gave their consent to take part in the study by completing and returning the questionnaire either via a web platform (round one) or by email (rounds two and three). Each round lasted six weeks (four weeks for participants to complete the questionnaire and two weeks for data analysis and preparation of the next questionnaire). Two weeks after the start of the round, participants who had

not yet responded were sent a reminder. Care was taken to ensure that no technical issues prevented participants from responding to the questionnaires.

2.4.1. Round one

We used the LimeSurvey Web platform (https://www.limesurvey.org) to conduct round one of the study. Participants received an e-mail invitation (sent via LimeSurvey) to take part in the study and were instructed to use the link provided to complete the online questionnaire.

In round one, participants were asked to express their opinion on the relevance (not important to very important) of 32 item-writing guidelines that related to either 1) content/subject/purpose of the clinical case (11 guidelines), 2) type of task required by the question (10 guidelines), 3) correction/scoring (5 guidelines), and 4) question design and development process (6 guidelines). We then asked participants to indicate the relevance of certain types of questions used for the assessment of clinical reasoning (6 types of questions). After each section, we asked if participants wanted to add elements (such as guidelines) or if they had any comment on the section. At the end of the questionnaire, we again asked if they had anything to add (e.g., guidelines, etc.) that was not mentioned in the preceding sections or any other comments. We also asked participants to indicate how the importance of given item-writing guidelines varied according to the context (student's level of education and assessment type) on a 10-point Likert scale (1 = not at all, to 10 = definitely). Finally, we gathered sociodemographic information about the experts. See Appendix C for an example of the round-one questionnaire.

We decided that the addition of suggested guidelines or question formats would be decided based on a consensus of all authors.

Descriptive statistics (median, mean, interquartile range (IQR), standard deviation, frequencies) were calculated for each of the guidelines and question formats in order to identify those to be included in the questionnaire for round two, as required by the Delphi method. The inclusion criteria were a median lower than 6 combined with an interquartile range less than or equal to two. Similar criteria have been used in other Delphi studies[21] and appeared appropriate for our context. During round one, we sought to identify guidelines on which an exclusion consensus could be reached (median lower than 6 combined with interquartile range lower than 2), that is, we aimed to identify guidelines that were not considered relevant for assessing clinical reasoning.

2.4.2. Round two

For this round, we chose to e-mail the participants a Word document containing personalized questionnaires, including their own individual answers from the previous round. An e-mail was sent to all participants from round one as well as to any "new experts" (experts identified by the participants 2 weeks after the start of data collection for round one).

Round two comprised 38 questions, broken down into the same five aforementionned sections as questionnaire 1 (4 sections for the 32 item-writing guidelines and 1 section for the six different types of questions used for assessing clinical reasoning). For each guideline and type of questions, the participants were given 1) the median of the experts' evaluations, 2) their own answer from the previous round, and 3) instructions that they could modify their responses (on the same 10-point Likert scale). Participants were also asked to tick off the contexts (preclinical, clerkship, post-doctoral, formative assessment, summative assessment) for each of the guidelines and question formats in which they would be relevant. Comments were gathered at the very end of the questionnaire instead of having specific sections for comments throughout the questionnaire. New participants received the same questionnaire, except they had no "previous answers." They also had to answer sociodemographic questions.

Descriptive statistics (median, mean, interquartile range, standard deviation, frequencies) were calculated for each guideline and question format in order to identify those to be included in the questionnaire for round three. The exclusion criterion was an interquartile range of less than or equal to 2. The objective of round two was to identify the guidelines and question formats on which there was already a consensus on relevance (or irrelevance), in order to survey round-three participants solely on the guidelines or question formats for which there was a consensus.

Moreover, if, for a particular context, a guideline or question format was deemed relevant by fewer than 30% of participants, we considered that there was a consensus on its irrelevance in that context. Conversely, if, for a particular context, a guideline or question format was deemed relevant by more than 70% of participants, we considered that there was a consensus on its relevance in that context. Given the results, in the case of overall relevance, only those guidelines or question formats for which a consensus was not achieved were held over in round three and, in the case of relevance in different contexts, only those guidelines or question formats touching on relevance in preclinical contexts for which a consensus was not achieved were held over.

2.4.3. Round three

For this round, we also chose to use a Word document sent by e-mail to participants. An e-mail was sent to all individuals who took part in at least one of the two first rounds. In the case of the guidelines or question formats for which no consensus was achieved, participants were given the opportunity to respond again after being informed of the median and interquartile range from round two. As a result, there were 12 guidelines in the first section of the third questionnaire, serving to assess, one last time, their relevance in assessing clinical reasoning.

In the case of relevance of guidelines or question formats at the preclinical level for which a consensus was not achieved, the participants only had to check off whether or not the guideline or question type was relevant for the context. The percentage of participants who deemed the guideline or question type to be relevant in the preclinical context was indicated. In this second section of the round-three questionnaire, participants had to assess the relevance of 18 guidelines and two question types relating specifically to the preclinical context.

The last section consisted of a single question asking participants to indicate the contexts in which it would be relevant to use written exams to assess clinical reasoning.

Descriptive statistics (median, mean, interquartile range, standard deviation, frequencies) were calculated for each of the guidelines and question types in order to identify the final selection of guidelines and question types deemed relevant and their context for use. At the end of round three, it was not deemed relevant to conduct a fourth round on the nine guidelines for which a consensus had not been previously achieved.

3. Results

The rate of response for each of the three rounds was 24% for round one (17/71; 62 participants initially identified + 9 new participants); 57% (12/21) for round two; and 62% (13/21) for round three. Twenty-one questionnaires were sent out in round two: 17 participants from round one along with the four experts newly proposed by round-one participants. For round three, only the individuals who took part in at least one of the first two rounds received the questionnaire. Four individuals took part in all three rounds; 10 participants took part in two rounds.

Table 1 provides the participants' characteristics, which only slightly varied from one round to the next.

Table 1
Characteristics of respondents for each round.

	Round one		Round two		Round three	
	Count	%	Count	%	Count	%
Language						
English	7	41.2	4	33.3	5	38.5
French	10	58.8	8	66.7	8	61.5
Category						
User	12	70.6	8	66.7	8	61.5
Researcher	5	29.4	2	16.7	3	23.1
User/Researcher	0	0	2	16.7	2	15.4
Experience						
Medical education	20.1 years	N/A	19.2 years	N/A	19.5 years	N/A
Assessing clinical reasoning	12.8 years	N/A	13.3 years	N/A	15.0 years	N/A
Occupation						
Family physician	4	23.5	4	33.3	3	23.1
Physician, other speciality	7	41.2	5	41.7	5	38.5
Other profession	6	35.3	3	25	5	38.5
If teacher, level primarily taught	11	64.7	9	75	9	69.2
Preclinical	5	N/A	4	N/A	4	N/A
Clerkship	6	N/A	6	N/A	5	N/A
Graduate	6	N/A	5	N/A	5	N/A
Undergraduate (disciplines other than medicine)	2	N/A	1	N/A	1	N/A
Master's (disciplines other than medicine)	2	N/A	2	N/A	2	N/A
Doctoral (disciplines other than medicine)	2	N/A	1	N/A	2	N/A
Level of expertise						
High level	4	23.5	4	33.3	4	30.8
Average level	3	17.6	1	8.3	2	15.4
Between high and average level	0	0	2	16.7	1	7.7
No opinion	10	58.8	5	41.7	6	46.2

It can be observed that most of the participants responded in French and identified themselves as teachers who used assessment instruments, rather than as researchers in the field of assessment. They had an average of about 20 years of experience in medical education and a little less than 15 years of experience in assessing clinical reasoning. They were nearly equally distributed between family medicine/medical specialties and the other professions. Most were involved in teaching and did so at different levels of the medical curriculum. A small number reported involvement in teaching activities in other programs, including graduate education. The majority of participants preferred not to comment on their level of expertise.

3.1. Round-one results

The response medians varied from 5 to 10, while the interquartile ranges varied from 1 to 6. The guidelines with a median of less than or equal to 5 (guidelines 3 and 20) had large interquartile ranges (IQR = 4). No guidelines or question types, therefore, were removed for the subsequent round. The median of the responses to the question *In your opinion, does the importance of the design guidelines presented above vary according to how far along students are in the curriculum (e.g., preclinical vs. clerkship)?* was 7. This justified verifying its relevance for each of the 32 guidelines and 6 question types when applied in the various contexts to be presented in the next round. Moreover, no respondent proposed adding a new guideline or question type to the questionnaire during round one. No guidelines were restated as the result of respondent comments.

3.2. Round-two results

As seen in Table 2, the results of round two had medians varying from 4.5 to 10 and interquartile ranges varying from 0 to 5. Consensus was achieved for 20 of the 32 guildelines, and all six question types. The medians of these 26 guidelines and question types were equal to or greater than 6, suggesting they were considered relevant by the panel of experts. As for the contexts for using the various rules for writing written-exam clinical reasoning assessment questions, it also appears that a consensus was achieved for nearly all the guidelines and question types when used for clerkship, post-doctoral medical education, and summative/formative assessment (guidelines deemed relevant by 70% or more of participants).

As shown in Table 2, compared to other guidelines, a greater number of participants deemed that the guidelines relating to Correction (i.e., Scoring) and Question Design Process (guidelines 22 to 26, and 27 to 32) were less relevant in a context of formative assessment (relevant for 42% to 67% of participants). There was no consensus on the use of guidelines in a preclinical context for 20 of the guidelines (guidelines deemed relevant by 30% to 69% of participants). Certain guidelines or question formats stood out because a consensus could not be reached on their irrelevance. This was the case for guidelines 2, 3, 7, 25, 21, and question format 35 at the preclinical level; guideline 17 at the post-doctoral level; and guideline 21 in the summative-assessment context.

3.3. Round-three results

In the case of the 12 guidelines presented in round three, the medians varied from 5 to 8 and the IQR varied from 1 to 3.75. These results indicate a consensus on the general relevance for nine additional guidelines (interquartile range less than or equal to 2). This led to a consensus being achieved for 28 of the 32 guidelines and 6 question formats (89%). No consensus was reached for guidelines 3, 21, 24, and 25. As for relevance for use in the preclinical context, there was no consensus after round three for seven guidelines (3, 14, 24, 25, 26, 27, and 29) and 1 question format (38) out of the 20 presented (see Table 2).

3.4. General relevance of the guidelines and question formats

With median values equal to or less than 6, there appears to be a consensus that guidelines 7, 15, and 20 were less relevant than the others in writing written-exam clinical reasoning assessment questions. Moreover, given their large interquartile range, denoting a lack of consensus among the experts, the responses for guidelines 3, 21, 24, and 25 suggest that these guidelines are also probably less relevant. With respect to context of use, the remaining guidelines all appear important for the clerkship, post-doctoral medical education, and summative assessment, with the exception of guideline 17 for post-doctoral medical education. Table 2 provides the list of guidelines and question formats as well as their relevance for assessment in each context. The table shows a greater variation in relevance in the preclinical context than for the other contexts.

4. Discussion

The objective of this study was to identify, through a consensus of experts, writing instructions that would make it possible to design good questions for assessing clinical reasoning. To this end, we used a modified Delphi study to collect the opinion of experts and to foster a consensus between them. Out of the 32 writing guidelines initially proposed, 25 were retained, three were rejected, and four did not achieve a consensus after three rounds. Most guidelines seemed important in the opinion of the experts, but four guidelines seemed more important (median $\geq$ 9 on ten-point scale): 1) The question is based on a clinical case, 2) Challenge is achievable for the student, 3) The correction scale (i.e.,

Table 2
List of retained guidelines and question formats, and their relevance according to four different contexts.

Item no.	Section name	Preclinical	Clerkship	Residency	Formative assessment	Summative assessment	Round two		Round three	
							Median	IQR	Median	IQR
Section A: Content / Subject / Purpose of the Clinical Case										
1	Based on a clinical case	x	x	x	x	x	9	1	N/A	N/A
2	Present raw data	**NR**	x	x	**NC**	x	8	2.5	8	1
3	Real patients	**NC**	**NC**	**NC**	**NC**	**NC**	5	3.75	5.5	2.5
4	Contextual information that a patient would have in all probability given during a medical interview	x	x	x	x	x	7.5	2	N/A	N/A
5	Negative information that a physician would look for	x	x	x	x	x	8	1.5	N/A	N/A
6	Varied length and level of detail, depending on the specific task	x	x	x	x	x	8	1.5	N/A	N/A
7	Complex clinical situation	**NR**	**NR**	**NR**	**NR**	**NR**	6	1	N/A	N/A
8	Realistic challenge given the students' level	x	x	x	x	x	10	0	N/A	N/A
9	Clinical situation with a degree of uncertainty	x	x	x	x	x	8	1.75	N/A	N/A
10	Asking a question about further information that a physician would look for in the specific situation in order to work towards a solution	x	x	x	x	x	8	2	N/A	N/A
11	Asking a question related to underlying assumptions or options pertinent to the situation	x	x	x	x	x	8	1	N/A	N/A
Section B: Type of Task Required by the Question										
12	Justify the posited diagnosis	x	x	x	x	**NC**	8	2.5	8	1.5
13	Make a clinical decision or take clinical action	x	x	x	x	x	8	3	8	1
14	Explain their clinical reasoning process	**NC**	x	x	x	**NC**	8	1.75	N/A	N/A
15	Take a stand with respect to the intervention's relevance, feasibility, and acceptability for the patient	**NR**	**NR**	**NR**	**NR**	**NR**	6	1.75	N/A	N/A
16	Demonstrate their clinical knowledge used	x	x	x	x	x	8	0	N/A	N/A
17	Demonstrate their fundamental knowledge used	x	x	**NR**	**NC**	**NC**	6.5	2.75	7	1.5
18	Summarize the information	x	x	x	x	x	8	0.75	N/A	N/A
19	Apply their clinical and fundamental knowledge beyond the learning context	x	**NC**	x	x	**NC**	8	2.5	8	1.75
20	Focus on a single aspect of the clinical reasoning process	**NR**	**NR**	**NR**	**NR**	**NR**	4.5	1.75	N/A	N/A
21	Demonstrate their complete clinical reasoning process	**NR**	**NR**	**NR**	**NR**	**NR**	7	3	7	2.5
Section C – Correction (i.e. Scoring protocol)										
22	Importance of the correction process (i.e., process of scoring and grading)	x	x	x	x	x	8	2	N/A	N/A
23	Grading scale provides the points for the various levels of response	x	x	x	**NC**	x	9	1	N/A	N/A
24	Scoring system involves weighting	**NR**	**NR**	**NR**	**NR**	**NR**	7	4	7	3

Table 2 (*continued*)

Item no.	Section name	Preclinical	Clerkship	Residency	Formative assessment	Summative assessment	Round two		Round three	
							Median	IQR	Median	IQR
25	Negative scores not be used with key-feature questions	**NR**	**NR**	**NR**	**NR**	**NR**	6	5	5	3.75
26	Score of 0 be given in the case of key-feature questions when a response involving dangerous or negligent behaviour is provided	**NC**	x	x	**NC**	x	8.5	2.75	8	2
Section D: Question Design Process										
27	Questions be designed by a panel of experts	x	x	x	**NC**	x	7	1	N/A	N/A
28	Questions be revised by a panel of experts	x	x	x	**NC**	x	9	1	N/A	N/A
29	Questions be pretested by a panel of experts	**NC**	x	x	**NC**	x	8	3	8	1.75
30	Correction key (i.e., marking key) be designed by a panel of experts	x	x	x	**NC**	x	8	3	8	2
31	Correction key (i.e., marking key) be revised by a panel of experts	x	x	x	**NC**	x	8	1.75	N/A	N/A
32	Correction key (i.e., marking key) be pretested by a panel of experts	x	x	x	**NC**	x	8	1.25	N/A	N/A
Section E: Types of Questions										
33	Multiple-choice question	x	x	**NC**	x	x	7	2	N/A	N/A
34	Short open-answer question	x	x	x	x	x	7	2	N/A	N/A
35	Script concordance test	**NR**	x	x	**NC**	**NC**	8	2	N/A	N/A
36	Extended matching questions	x	x	x	x	x	8	1	N/A	N/A
37	Modified essay questions	x	**NC**	x	**NC**	x	7	1.5	N/A	N/A
38	Key-feature questions	**NC**	x	x	x	x	8	1	N/A	N/A

Notes. x: Consensus (at least 69% of respondents agreed with the item); NC: No consensus; NR: Not relevant; N/A: Not applicable; IQR: Interquartile range

scoring grid) is explicit, and 4) A panel of experts revises the questions. No participant proposed new guidelines for writing clinical reasoning assessment questions. These results may suggest that the initial list of item-writing guidelines used for the round-one survey was comprehensive.

The participants appear to be from varied settings and contexts, which may have contributed to the difficulty in reaching a consensus on the importance of guidelines and question formats, and especially with respect to the relevant contexts for their use. The participants' varied backgrounds and the type of medical curriculum in which they work may have influenced their responses. In particular, we think that traditionally designed curricula,[30] whose first years are devoted mostly to the acquisition of theoretical knowledge, leave little room for learning clinical reasoning and, consequently, for assessing it. Even if there is a recent trend in teaching and assessing clinical reasoning at the preclinical phase of the medical curriculum,[8] it may have been more difficult for our experts to imagine which guidelines would be the most important to apply in order to assess clinical reasoning during the preclinical phase of the medical curriculum, since their "expert" responses were not dictated by their practical experience. This might reflect actual practices and should not necessarily be interpreted as a perception that assessing clinical reasoning is not as relevant during the preclinical phase of medical curricula as it would be later in the learning process.

In addition, we cannot overlook the fact that there are many conceptualizations of clinical reasoning, as recently illustrated in work by Young et al.[31] Having little consensus between experts with such different backgrounds may be an illustration of these underlying differences.

Several guidelines (3, 21, 24, and 25) were not maintained in our final list given the persistent variation in the responses of experts even after three rounds. Moreover, the medians for two of these four guidelines were also less than or equal to 6 (guidelines 3 and 25), which suggests that these guidelines would have been rejected in a subsequent round once a consensus had been reached. Guideline 3, which was eliminated due to lack of consensus, also appeared to have a lack of consensus as to its relevancy depending on the context; and guidelines 21, 24, and 25 were deemed not relevant in all contexts. This might account in part for the difficulty in achieving a consensus on their overall importance. Similar non-relevance in all contexts was observed for guidelines 7, 15, and 20, which might explain why they were also rejected. Other guidelines for which relevance seemed to have varied depending on the context are the ones related to question design process (6 guidelines), which were deemed less important in formative assessment. This lack of consensus makes sense since the question design process may be less formal for formative assessment than for summative assessment for some people, whereas others may think it requires the same rigor.

Finally, we observed that participants were unable to state an opinion on whether the task given to students had to allow them to demonstrate their complete clinical-reasoning process (guideline 21), whereas they indicated that it isn't important for the task to necessarily focus on a single aspect of clinical reasoning (guideline 20). The most plausible explanation for this finding is that, in certain contexts, it may be relevant to ask students to focus on a single aspect and, in other contexts, the complete demonstration of the clinical-reasoning process is important. This may reflect the multiple conceptualizations about clinical reasoning that have, since the data collection phase of this study, been identified in the literature.[31]

The six types of exam questions (MCQ, OEQ, etc.) all appeared to be equally relevant to the participants in assessing clinical reasoning. This result is in line with the idea that it is not the question format that matters, but the way the question is written.[14,18]

This study has some limitations. The participation rate proved lower than anticipated. It is unfortunate that only four participants took part in all three rounds, since this limits the significance of the study results. We are unable to state whether those who didn't respond had profiles that differed from the participants and even less able to tell if obtaining their opinions about the subject would have modified the results obtained. Using another collection method, such as the nominal group, might have been more conducive to broader participation and respondent stability throughout the study. Nevertheless, these methods require bringing individuals together in the same location, which is an obstacle circumvented by the Delphi method, allowing a large number of experts to be consulted without having to be together in the same room. The experts contacted for the study constitute a convenience sample. Nevertheless, the multiple approaches used for recruitment should have minimized bias impact. We were surprised that the majority of participants didn't view themselves as experts on the subject. A great deal of care went into formulating the guidelines, and the questionnaire was pretested. It is possible that participants may have interpreted guidelines in a manner other than what we

intended. In particular, the term "clinical reasoning" was not defined in the questionnaire. However, since consensus was reached on nearly all the guidelines and question formats, it makes it more unlikely that any misunderstanding of terms would have led to bias. The literature review conducted to generate the initial list of guidelines to include in the round-one questionnaire came from another study. No participant suggested including additional guidelines in the next rounds, so it therefore seems unlikely that an important concept was not included.

5. Conclusions

Assessing clinical reasoning from the beginning of the medical curriculum, represents a major cultural change in some faculties where new competency-based curriculum are being implemented. Tools are needed to help teachers modify their assessment practices accordingly. Despite a low participation rate in this study, participants arrived at a consensus in the identification of a large number of guidelines as being relevant. These results suggest that designing written-exam clinical reasoning assessment questions is indeed a major challenge. Four guidelines were highly recommended by the respondents. As a future step, the remaining guidelines will be grouped into categories and combined with the four highlighted guidelines in order to ensure a more accessible and user-friendly format (tool). This tool would then need to be validated. Future studies will investigate the appropriateness of this tool when used to generate written questions that effectively assess clinical reasoning.

Acknowledgements

We would to thank Kathleen Ouellet and Marianne Xhignesse for their critical review of the manuscript.

Funding sources

This work was supported by the Société des médecins de l'Université de Sherbrooke, Fonds de développement pédagogique 2013.

Appendix A. Supplementary material

Supplementary data associated with this article can be found in the online version at https://doi.org/10.1016/j.hpe.2018.09.001.

References

1. Norman GR. Research in clinical reasoning: past history and current trends. *Med Educ* 2005;39(4):418–427 http://dx.doi.org/10.1111/j.1365-2929.2005.02127.x.
2. Schmidt HG, Boshuizen HPA. On acquiring expertise in medicine. *Educ Psychol Rev* 1993;5(3):205–221 http://dx.doi.org/10.1007/BF01323044.
3. Durning S, Dong T, Artino A, van der Vleuten C, Holmboe E, Schuwirth L. Dual processing theory and experts' reasoning: exploring thinking on national multiple-choice questions. *Perspect Med Educ* 2015;4(4):168–175 http://dx.doi.org/10.1007/s40037-015-0196-6.
4. Norman G, Sherbino J, Dore, K, et al. The etiology of diagnostic errors: a controlled trial of system 1 versus system 2 reasoning. *Acad Med* 2014;89(2):277–284 http://dx.doi.org/10.1097/ACM.0000000000000105.
5. Young M, Thomas A, Lubarsky, S, et al. Drawing boundaries: the difficulty in defining clinical reasoning. *Acad Med* 2018.
6. Higgs J, Jones M, Loftus S, Christensen N. *Clinical Reasoning in the Health Professions*, 3rd ed., Sydney, AUS: Elsevier; 2008.
7. Rencic J. Twelve tips for teaching expertise in clinical reasoning. *Med Teach* 2011;33(11):887–892.
8. Gay S, Bartlett M, McKinley R. Teaching clinical reasoning to medical students. *Clin Teach* 2013;10:308–312.
9. Schmidt HG, Mamede S. How to improve the teaching of clinical reasoning: a narrative review and a proposal. *Med Educ* 2015;49(10):961–973 http://dx.doi.org/10.1111/medu.12775.
10. Askew K, Manthey D, Mahler S. Clinical reasoning: are we testing what are teaching?. *Med Educ* 2012;46:540–542.
11. van der Vleuten C, Norman G, Schuwirth L. Clinical reasoning in the health professions. In: HIggs J, Jones MA, Loftus S, NC, editors. *Assessing Clinical Reasoning*, 3rd ed., Sydney: Elsevier; 2008.
12. Boileau E, St-Onge C, Audetat M. Is there a way for clinical teachers to assist struggling learners? A synthetic review of the literature. *Adv Med Educ Pract* 2017;8:89–97 http://dx.doi.org/10.2147/AMEP.S123410.
13. Case SM, Swanson DB, Ripkey DR. Comparison of items in five-option and extended-matching formats for assessment of diagnostic skills. *Acad Med* 1994;69(10 Suppl):S1–S3.
14. Coderre SP, Harasym P, Mandin H, Fick G. The impact of two multiple-choice question formats on the problem-solving strategies used by novices and experts. *BMC Med Educ* 2004;4:23 http://dx.doi.org/10.1186/1472-6920-4-23.
15. Page G, Bordage G. The Medical Council of Canada's key features project: a more valid written examination of clinical decision-making skills. *Acad Med* 1995;70(2):104–110.
16. 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(4):189–195 http://dx.doi.org/10.1207/S15328015TLM1204_5.
17. Palmer EJ, Devitt PG. Assessment of higher order cognitive skills in undergraduate education: modified essay or multiple choice questions? Research paper. *BMC Med Educ* 2007;7(1):49.
18. Beullens J, Struyf E, Van Damme B. Do extended matching multiple-choice questions measure clinical reasoning?. *Med Educ* 2005;39(4):410–417.
19. Haladyna TM, Downing SM, Rodriguez MC. A review of multiple-choice item-writing guidelines for classroom assessment. *Appl Meas Educ* 2002;15(3):309–334.
20. Okoli C, Pawlowski SD. The Delphi method as a research tool: an example, design considerations and applications. *Inf Manag* 2004;42(1):15–29.

21. von der Gracht H. Consensus measurement in Delphi studies: review and implications for future quality assurance. *Technol Forecast Soc Change* 2012;79:1525–1536.
22. Vachon-Lachiver E, St-Onge C, Cloutier J, Farand P. L'identification de consignes discriminantes pour la rédaction d'examens écrits. *Pédagogie Médicale* 2018;18: 55–64.
23. Haladyna TM, Downing SM. Validity of a taxonomy of multiple-choice item-writing rules. *Appl Meas Educ* 1989;2: 51–78 http://dx.doi.org/10.1207/s15324818ame0201_3.
24. Case S, Swanson D. *Constructing Written Test Questions for the Basic and Clinical Sciences*. Philadelphia, PA: National Board of Medical Examiners; 1998.
25. Schuwirth LWT. How to write short cases for assessing problem-solving skills. *Med Teach* 1999;21:144–150.
26. Auger R, Séguin SP, Nézet-Séguin C. *Formation de base en évaluation des apprentissages: la planification de la mesure des apprentissages dans une démarche d'évaluation*. *Outremont*. Canada: Éditions Logiques; 2000.
27. Farmer EA, Page G. A practical guide to assessing clinical decision-making skills using the key features approach. *Med Educ* 2005;39(1188):1194.
28. Frey BB, Petersen S, Edwards LM, Pedrotti JT, Peyton V. Item-writing rules: collective wisdom. *Teach Teach Educ* 2005;21: 357–364.
29. Fournier J, Demeester A, Charlin B. Script concordance tests: guidelines for construction. *BMC* 2008;8:18.
30. Irby DM, Cooke M, O'Brien BC. Calls for reform of medical education by the Carnegie Foundation for the Advancement of Teaching: 1910 and 2010. *Acad Med* 2010;85(2): 220–227 http://dx.doi.org/10.1097/ACM.0b013e3181c88449.
31. Young M, Dory V, Lubarsky S, Thomas A. How different theories of clinical reasoning influence teaching and assessment. *Acad Med* 2018;93(9):1415 http://dx.doi.org/10.1097/ACM.0000000000002303.