Electronic assessment of clinical reasoning in clerkships

MEDICAL TEACHER, 2017
VOL. 39, NO. 5, 476–485
http://dx.doi.org/10.1080/0142159X.2017.1297525

MEDICAL TEACHER
Taylor & Francis
Taylor & Francis Group

Electronic assessment of clinical reasoning in clerkships: A mixed-methods comparison of long-menu key-feature problems with context-rich single best answer questions

Sören Huwendiek[a], Friedrich Reichert[b], Cecilia Duncker[c], Bas A. de Leng[d], Cees P. M. van der Vleuten[e], Arno M. M. Muijtjens[e], Hans-Martin Bosse[f], Martin Haag[g], Georg F. Hoffmann[h], Burkhard Tönshoff[h] and Diana Dolmans[e]

[a]Department of Assessment and Evaluation, Institute of Medical Education Bern, University of Bern, Bern, Switzerland; [b]Department of Pediatric Cardiology and Intensive Care Medicine, Klinikum Stuttgart, Stuttgart, Germany; [c]Clinic for Child and Adolescent Psychiatry, University Hospital Kiel, Germany; [d]Institute of Medical Education (IfAS), Faculty of Medicine, University of Muenster, Münster, Germany; [e]Department of Educational Development and Research, Maastricht University, Maastricht, the Netherlands; [f]Clinic for General Paediatrics, Neonatology and Paediatric Cardiology, University Children's Hospital Düsseldorf, Düsseldorf, Germany; [g]GECKO Institute of Medicine, Informatics & Economics, Heilbronn University, Heilbronn, Germany; [h]Clinic I, University Children's Hospital Heidelberg, Heidelberg, Germany

ABSTRACT

Background: It remains unclear which item format would best suit the assessment of clinical reasoning: context-rich single best answer questions (crSBAs) or key-feature problems (KFPs). This study compared KFPs and crSBAs with respect to students' acceptance, their educational impact, and psychometric characteristics when used in a summative end-of-clinical-clerkship pediatric exam.

Methods: Fifth-year medical students ($n = 377$) took a computer-based exam that included 6–9 KFPs and 9–20 crSBAs which assessed their clinical reasoning skills, in addition to an objective structured clinical exam (OSCE) that assessed their clinical skills. Each KFP consisted of a case vignette and three key features using a "long-menu" question format. We explored students' perceptions of the KFPs and crSBAs in eight focus groups and analyzed statistical data of 11 exams.

Results: Compared to crSBAs, KFPs were perceived as more realistic and difficult, providing a greater stimulus for the intense study of clinical reasoning, and were generally well accepted. The statistical analysis revealed no difference in difficulty, but KFPs resulted more reliable and efficient than crSBAs. The correlation between the two formats was high, while KFPs correlated more closely with the OSCE score.

Conclusions: KFPs with long-menu exams seem to bring about a positive educational effect without psychometric drawbacks.

Introduction

Various question formats have been described for the assessment of clinical reasoning (Higgs et al. 2008). The impact of each of these formats on student learning is still not well understood. Better understanding of this "pre-assessment effect" (Cilliers et al. 2012) would pave the way for clinical clerkship directors to better steer student learning through the concluding assessment. We therefore conducted a study comparing two different item formats in terms of their impact on student learning and their relevant psychometric characteristics.

When comparing item formats several aspects are important. A test item essentially consists of two parts, that is, the stimulus and the response part (Schuwirth & van der Vleuten 2004). Whereas the former refers to the task imposed by the stem of an item, e.g. a case vignette, the latter denotes the method that examinees use to indicate their responses (Schuwirth & van der Vleuten 2004). The stimulus format can be either context-free or context-rich and shape the focus of the question (Schuwirth & van der Vleuten 2004). While context-free stimuli usually measure factual knowledge, context-rich stimuli, by contrast, serve to assess applied knowledge by presenting a specific scenario and asking for decisions, focusing on key features to solve a clinical problem, for example (as when a case vignette is used). The comparison we make in the present study is between two question formats with a *context-rich* stem designed to assess clinical reasoning.

Practice points

- The analysis of focus group discussions revealed that students perceived KFPs with long-menu questions as providing a greater stimulus for the intense study of clinical reasoning than did crSBAs.
- Statistically, KFPs revealed a higher efficiency than crSBAs.
- This study supports the idea that, from an educational perspective, both the stimulus and response format of questions are important.
- Including KFPs with long menu in clerkship examinations seems to offer valuable opportunities to steer learning in clinical clerkships without psychometric drawbacks.

CONTACT Sören Huwendiek soeren.huwendiek@iml.unibe.ch Department of Assessment and Evaluation, Institute of Medical Education, Faculty of Medicine, University of Bern, Konsumstrasse 13, CH-3010 Bern, Switzerland

As regards the response part, this can be grouped into two broadly defined categories: multiple choice-type questions (e.g. single best answer and multiple true/false questions) and open-ended (e.g. write-in) questions (Schuwirth & van der Vleuten 2004). Schuwirth et al. (1996) demonstrated that electronic long-menu questions are an equivalent alternative to open-ended questions in computerized assessment. Long menus are alphabetically ordered long lists of (over 500) possible answers that prevent a cueing effect, because one has to type the solution into a dialog field (Schuwirth et al. 1996). The computer then searches through the long-menu list for "hits". The alternatives found are immediately presented to the examinee, so he or she can check whether the retrieved option is the desired one.

Besides varying in question types, context-rich items may also differ in the number of questions they contain: a case vignette for instance, may be followed by a set of multiple questions. One approach that is widely used to assess clinical reasoning, and that usually groups several questions together, is to use key features (Bordage et al. 1995; Page & Bordage 1995). A key feature is defined as a critical step in the resolution of a problem, one where examinees are most likely to make a slip when trying to resolve the problem or one that complicates the identification and management of the problem in practice (Page & Bordage 1995). Problems comprising a key feature, referred to as key-feature problems (KFPs), consist of a brief stem with a short patient vignette (stimulus format) containing relevant and non-relevant elements, such as symptoms and findings, followed by one or more questions. KFP-based examinations, as such, do not follow a fixed item format, but should rather be seen as an *approach* to testing.

Studies on KFPs, however, have demonstrated that assessments yielded the best psychometric characteristics when responses were recorded in short-menu, write-in or electronic long-menu format and vignettes were followed by two to three questions each (Bordage et al. 1995; Page & Bordage 1995, Fischer et al. 2005; Norman et al. 2006; Hrynchak et al. 2014; Bronander et al. 2015). In a comprehensive review of the literature regarding the reliability and validity of KFPs, Hrynchak et al. (2014) conclude that published research supports the use of KFP-based examinations for the assessment of clinical reasoning. Their internal consistency reliability, as measured by Cronbach's alpha, has generally been reported to be acceptable. Yet, the review acknowledges that all categories of validity evidence should be subjected to further scrutiny, as outcomes are contingent upon many contextual factors, including the population to which the test is administered, following which validity cannot be considered as a universal function of an assessment format. Considering the foregoing, and the fact that, to our knowledge, assessment at the *clerkship* level has received scant attention (Hatala & Norman 2002), we believe it important to delve deeper into this subject matter. Hatala and Norman (2002) introduced the KF assessment method into an *undergraduate* setting when evaluating the clinical decision-making skills of internal medicine clerkship students through a 2-h paper-based exam consisting of 15 KFPs. They reported a Cronbach's Alpha of 0.49.

Another way to assess clinical reasoning is by means of context-rich single best answer (crSBA) multiple choice-type questions with case vignettes as stimulus format (Higgs et al. 2008). In such format the vignette usually is followed by only one question (Higgs et al. 2008), often containing 3–5 response options with only one correct answer (single best answer format, A-type questions).

At present, there is a paucity of evidence on the educational effects such different item formats have. It is widely anticipated that whatever item format is used, the stimulus format would be more important in determining what is tested than the response format (Schuwirth & van der Vleuten 2004). As single best answer question types are considered more beneficial in terms of their effectiveness, grading system, accountability and costs involved, these are often preferred to open-ended questions such as write-in or electronic long-menu questions (Elstein 1993; Downing 2002, 2009; Desjardins et al. 2014). A few studies, however, have pointed out that the response format may also play a role, as cueing may influence item difficulty, with open-ended questions being more difficult than closed-ended questions (Heemskerk et al. 2008; Desjardins et al. 2014). These perceptions potentially influence the educational effect of an exam, making it interesting to investigate whether this is indeed the case, and if so, what this effect will be.

To our knowledge, no study has ever investigated the educational effect of using KFPs or crSBAs for the assessment of clinical reasoning in clinical clerkships. As it is generally accepted that assessment drives learning (McLachlan 2006; Cilliers et al. 2012; van der Vleuten & Schuwirth 2005), we felt it would be valuable to investigate students' perceptions of these specific item formats which both use a context-rich stimulus but differ in terms of the response format (open-ended versus closed-ended) and to investigate whether these perceptions are also mirrored in the psychometric characteristics. Hence, the purpose of this study was to investigate students' perceptions of KFPs and crSBAs and to compare both formats in terms of their impact on student learning and psychometric characteristics when used as a summative end-of-clinical-clerkship exam. In medical undergraduate education, objective structured clinical exams (OSCE, Harden et al. 1975) are widely used to test several competencies, including practical skills, communication skills and clinical reasoning, in a simulated setting, often using standardized patients. This form of assessment is generally judged as quite realistic and its tasks demand active knowledge, as possible cues (such as predefined answers in Type-A questions) are not present. We therefore consider it meaningful to set the results of the two question formats against those of the OSCE to see whether any one format (the one that requires more active knowledge, for example) correlates more closely with the OSCE results. Consistent with the aim we formulated the following research questions:

a. How do students perceive KFPs with an electronic long-menu response format and crSBAs used for the summative assessment of their clinical clerkship, particularly when it comes to the (educational or pre-assessment) effect on their learning and how do they accept both formats?
b. How difficult and reliable (*and* efficient) are KFPs compared to crSBAs? Do the results of these two formats correlate and how do these correlate with the total OSCE score in the respective end-of-clerkship exams?

Methods

Context and participants

Participants were fifth-year medical students ($n = 377$) from Heidelberg Medical School doing regular pediatric clerkships, which included 10 virtual patients (VPs) that were integrated into the curriculum (Huwendiek et al. 2013) and specifically designed (Huwendiek et al. 2009) to foster clinical decision-making. Each group of students taking the same exam also included some medical exchange students from abroad. At the end of their clerkship, students were subjected to a 1-h computer-based exam (Huwendiek et al. 2007), consisting of KFPs, context-rich and context-free SBAs. As context-free SBAs target factual knowledge rather than clinical reasoning (Schuwirth & van der Vleuten 2004), we did not include these questions in our study. Additionally, the end-of-clerkship assessment included an 11-station pediatric OSCE (5 min per station, overall testing time also approximately 60 min).

Instruments

We used the CAMPUS assessment software as assessment tool (Heid et al. 2006). This software tool facilitates the incorporation of many different question formats and has been used in the regular assessment of medical students for more than 15 years.

Key-feature problems

Cases were constructed according to the key-feature approach (Page & Bordage 1995) and followed published recommendations (Farmer & Page 2005; Kopp et al. 2006). Each KFP consisted of a case vignette (context-rich stem) and three key features (KF) using "long menu" as response format requiring free-text entry (Schuwirth et al. 1996). In our study, students selected an answer from the list by typing it into a dialog box. The computer then searched through the long-menu list for "hits". The alternatives found were immediately presented to the examinee, so he or she could check whether the retrieved option was the desired one. An additional free-text comment field was available in case students did not find a suitable answer in the long menu. To encourage students to use the long menu, we informed them that they would earn only half of the points if they used the free-text comment field and that the long menu contained the correct answer. The electronic system forced students to first answer question 1 before moving on to question 2 or 3. They had the possibility to go back and read (e.g. to have a look at the vignette again), but were no longer able to change their answer.

crSBAs

These questions consisted of a case vignette (context-rich stem) and focused on one relevant key feature. To answer each question, students had to select one correct answer from among a list of five options (A-type question).

For examples of the KFP and crSBA formats, see Appendix 1.

Blueprint

The Heidelberg catalog of learning objectives (Bosse et al. 2011) served as the blueprint for the assessment questions, with certain percentages of questions from relevant domains (e.g. pediatric infectious diseases, pediatric oncology and hematology, pediatric radiology, pediatric cardiology, general pediatrics, pediatric rheumatology, and pediatric pulmonology) to ensure the exam covered a broad range of topics. KFPs mainly assessed eight important pediatric leading symptoms including "fever without focus", "coughing", "vomiting", "diarrhea", "limping", "abnormal appearance of urine", "skin alteration", and "edema". We specifically chose a KFP-based item format with long menu, as we hypothesized that the open response format would foster student learning more than a closed-question format like crSBA would. The crSBAs mainly assessed pediatric clinical topics of the Heidelberg catalog of learning objectives that were different from the eight leading symptoms.

Review

Assessment questions were reviewed by at least two pediatricians and in many instances they were also subjected to scrutiny by a regional pediatric assessment alliance that regularly revised and exchanged those (Walter et al. 2008).

OSCE

The OSCE consisted of 10–11 pediatric stations assessing clinical skills (5 min per station, overall testing time of 60 min). The blueprint was comprised of three stations that focused on history-taking and communication skills with standardized patients playing the role of "mothers" and "fathers", three stations targeting the physical exam with mannequins, one to two media stations (e.g. characterizing a neurological fit, facial paralysis), a percentile station, radiograph/sonographic interpretation station and procedural skills station (e.g. resuscitation, bladder puncture, and lumbar puncture) (Bosse et al. 2006).

We investigated the research questions using Fokus groups and statistical methods. For visualization of the methods used please see Figure 1.

Focus group study

Three groups of approximately 40 clerkship students ($n = 116$ in total) received an invitation to participate in a study on clinical education. Focus group interviews were held after students had completed both the rotation and the exam. To avoid bias, only those students who consented to participate received additional information about the study. Eventually, we randomly selected 39 students out of a total of the 65 students (56%) who had indicated their interest in the study and divided them into eight focus groups of approximately 4–7 students. Participants were offered a small compensation. During the interviews, we also explored other clerkship-related aspects the results of which have been reported elsewhere (Huwendiek et al. 2013).

Prior to analysis, the interviews were videotaped and transcribed verbatim. We selected a qualitative focus group

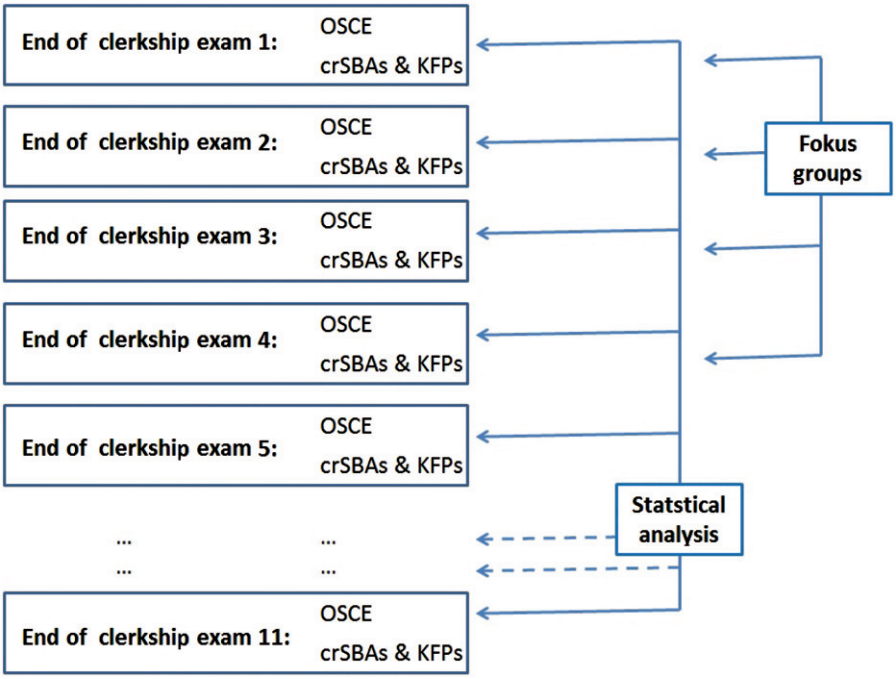

Figure 1. Flow chart of methods used.

methodology, as it has the potential to bring into focus not only the perceptions of participants but also the ideas and contemplations behind them (Krueger & Casey 2000). Interviewing peers in groups rather than individually fosters a safe environment because it decreases the power distance between researcher and subjects and encourages in-depth discussion, which, in turn, may ultimately cause participants to change or adjust their views. To reduce any social pressure within the group that could prevent participants from speaking out freely, the moderator made clear that contributions of any kind would be highly valued and that each opinion or view, no matter how divergent, would be respected. Prior to the discussion, participants were expected to write down their ideas about the theme under discussion. The lead author (Sören Huwendiek) who is well versed in moderating focus group discussions conducted all the interviews which lasted two hours each and took place on different days. Group interviews were guided by a questioning route (Appendix 2) which promoted consistency between interviews (Krueger & Casey 2000). Two researchers (Cecilia Duncker, Friedrich Reichert) assisted the moderator by taking detailed notes and recording the interviews on video. We transcribed the tapes verbatim and sent summaries of the results of the respective focus groups to participants for approval.

Three authors (Cecilia Duncker, Friedrich Reichert, and Sören Huwendiek) subjected the scripts to a qualitative thematic analysis. This process consisted of a careful reading of the scripts, during which sentences or fragments of sentences were identified as a code (Braun & Clarke 2006). From among these codes, themes and subthemes were identified and subsequently discussed in the research team in an iterative process, which elaborations served as new input for the coding process. These steps were repeated until all researchers were in agreement. Whenever possible, we kept an account of the strength of opinions and the number of times group members shared similar views. We ended the interviews after the eighth session, because the latter interview did not deliver us any new insights (Krueger & Casey 2000).

Statistical analysis

We used the questions that assessed clinical reasoning (KFPs and crSBAs) of 11 end-of-clerkship summative exams (377 fifth-year medical students) as input for the statistical analysis. The exams all had the same blueprint but differed in the questions used. In line with recommendations by Page and Bordage (1995), we considered each case in the KFP component an item, and used a partial scoring system (a score between 0 and 1 reflecting the proportion of correct responses). We analyzed difficulty (percentage of correct answers) at the item level for the pool of items over 11 exams and compared the difficulty distributions for the two question formats (KFPs and crSBAs). In order to investigate the reliability of the mean KFP score, we performed a generalizability analysis. For each exam we estimated variance components using a simple all-random person-by-item design. We then pooled variance components across exams weighted by sample size. The pooled variance components were used to estimate the generalizability coefficient (G coefficient), which is conceptually similar to Cronbach's alpha. The same procedure was applied to obtain the G coefficients of the crSBA questions. We were interested in comparing the reliabilities of KFPs and crSBAs that take equal amounts of testing time. To this end we also computed the G coefficients of KFPs and crSBAs for varying hypothetical numbers of items, hence varying amounts of testing time, according to the equation

$$G = \frac{V_p}{V_p + V_{pi}/N_i}$$

where V_p is the person variance, the variance of interest, V_{pi} is the person-item variance, the error variance, and N_i is the number of items (Brennan 2001). In addition, efficiency of the two question formats was calculated using the Spearman–Brown prediction formula (Norman et al. 1996).

Test duration was determined by using the log files from the server data bank of the assessment system. To estimate how long students needed to answer a question, we measured the lapse of time from the moment students opened a page to a question until they moved on to the next page (to the next question). Since KFPs consisted of three questions, we added the times of the respective three questions. If students went back to a specific question, we added all times they returned to this question. Times were added even when an answer remained unchanged.

We calculated Pearson's correlation for student's component scores for each of the 11 exams (exam level, student-level data). In order to reduce bias when estimating the average correlation, the scale of the correlations was Fisher-z transformed before calculating the average. To obtain true correlations we divided the observed correlations by the square root of the product of both reliabilities. Finally, we computed correlations of the two different written assessment formats with the total OSCE score. True correlations obtained after correction for attenuation should be interpreted with caution, because in the case of low reliability in particular, there may be a considerable risk of over-correction (Muchinsky 1996).

Ethical approval

In Germany, where the present study was carried out, this type of educational study does not require approval from an ethics committee. Nevertheless, we confirm that participants took part on a voluntary basis, cannot be identified by the material presented and run no conceivable risk by having taken part in the study.

Results

Focus group study

During the analysis, four main themes emerged, which are detailed below. Table 1 lists quotes from the interviews that are illustrative of each theme with the numbers in parentheses referring to the focus group and student number, respectively. Also see Figure 2 for a visual summary of the results.

Table 1. Quotes from focus groups: numbers in parentheses refer to focus group and student number, respectively.

(i) KFPs were perceived as more realistic because of the long-menu questions
"KFPs are clearly more realistic, as you have to choose the correct answer yourself like in reality, and not from a predefined list". (1/2)
"In the KFP you must come up with an answer yourself, if you do not have one, you have a problem, which is also the case in real clinical practice". (5/6)
"If you are in an emergency situation, you do not have time to look up the correct therapy either, so I find these long menus more realistic than crSBAs". (3/1)
(ii) KFPs were perceived as more difficult to answer because of the long-menu questions
"As you have to know everything in the long-menu questions actively by heart, it is more difficult. The crSBAs are easier to solve as you will more easily recognize the correct answer from the options listed". (6/6)
"These long-menu questions were more difficult as you really had to reproduce your knowledge which is different from just having to recognize the correct answer". (5/4)
"When I had to give a diagnosis in a KFP, I found it much harder because I had to write it myself. In the crSBA I could see the options, which was easier". (4/1)
"When you enter your answer in the long menu it is harder to judge how well you have done in this question than in the crSBA questions". (1/4)
(iii) KFPs provided a greater stimulus for the intense study of clinical decision-making than did crSBAs
"As you knew that the eight leading symptoms would be assessed with long-menu questions, you studied these leading symptoms much harder and more actively and thought about what they might ask you in the exam to be well prepared". (1/2)
"As I knew that eight leading symptoms would be tested with long menu and I had to write it down myself in the exam, I studied these issues much more intensely, e.g. by working though the VP more carefully and repeatedly and taking notes while working through the cases". (7/4)
"I studied more for the long-menu questions. Had I known the exam was purely crSBA-based, I would have studied less, just the lecture notes, e.g. and I would have studied the main aspects of them, because this often suffices when answering multiple choice questions for which you only need to tick the correct answer". (5/5)
"Besides working through the VP several times I also explicitly wrote small guidelines in which I summarized for myself differential diagnoses that are typical of a leading symptom, typical diagnostic procedures and therapeutic options. Since I knew this would be relevant for the exam with KFPs I did that". (4/2)
"Long-menu questions in the VP or KFP much better show you your gaps in knowledge. This motivated me. In the crSBA it is easier to pretend that you know". (2/4)
"Learning with VPs was the best way to prepare for the KFPs, as you actively practiced how to think and proceed in such a case. They taught me more than the problem-based learning tutorials did". (1/5)
(iv) Overall, KFPs were received positively and perceived to support the learning of clinically relevant topics; however, some aspects need to be taken into account when using them for high-stakes examinations.
"The KFPs were not so easy but I appreciate them better than crSBAs because they motivate you to learn relevant things which you will need in any case for your studies and future practice". (1/1)
"I liked the VPs and KFPs as I knew these would help me learn and test my knowledge of relevant topics, which is important to your future professional life, knowing how to proceed in cases of a patient with such a leading symptom". (4/4)
"I appreciated the KFPs as they showed you what you really knew and had understood and could apply and what not". (3/5)
"Sometimes I could not find the word I had in mind in the long menu—that was frustrating. At that point the additional free-text field was helpful". (1/4)
"I think it would be helpful if you could always give a comment on the item you have chosen from the long menu including why you think something is important. Similar to an oral exam … ". (1/2)

KFPs were perceived as more realistic because of the long-menu questions

Students perceived the KFPs with long-menu questions as more realistic than the crSBAs, as these required them to actively produce the solution as is the case in real life when there is no opportunity to choose from options.

KFPs were perceived as more difficult to answer because of the long-menu questions

Students perceived the long-menu questions as more difficult to answer as there was no limited range of options to choose from. The long-menu response format required students to generate the answer themselves. Overall, students found the crSBA format to assess rather passive knowledge and the long-menu response format of KFPs to assess more active knowledge. Moreover, students had more difficulty judging how well they performed in the long-menu questions compared to the crSBAs.

KFPs provided a greater stimulus for the intense study of clinical decision-making than did crSBAs

As the KFP with long-menu questions were perceived as more difficult and more realistic than crSBAs, students were

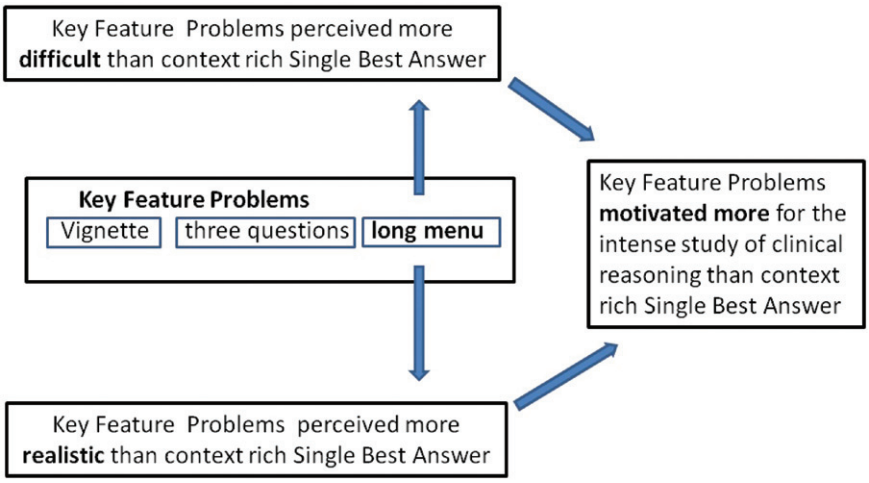

Figure 2. Visualization of results of the focus group study. The arrows represent the influence. E.g. the long-menu response format of KFPs had an influence on the overall perceived realism and difficulty of KFPs.

more motivated to study hard and more in depth for the KFP assessment questions. Likewise, students reported KFPs motivated them more to use VPs for learning. More specifically, they worked through VPs several times and very ambitiously as they knew that these might be helpful in answering KFP questions. VPs were perceived as the best way to prepare for the KFPs with long-menu questions, whereas problem-based learning, seminars and learning from books were regarded as less helpful. Students found that actively working through the VPs (with embedded long-menu self-assessment options), including the helpful feedback they provided, prepared them well for the KFPs. Students generally felt that VPs afforded the best opportunities to actively train the clinical reasoning competencies required for solving the KFPs. Without these VPs, it would have been much harder to learn about clinical reasoning for the exam.

Overall, KFPs were received positively and perceived to support the learning of clinically relevant topics; however, some aspects need to be taken into account when using them for high-stakes examinations

In general, KFPs were favorably received by students: They liked the use of KFPs for assessment purposes, for the incentive these provided them to learn important topics that were clinically and practically relevant. It is important that the content of the long menus be exhaustive and includes all synonyms relevant to each specific question. Omission of any synonyms caused frustration among students if they could not find the correct answer and had to produce synonyms themselves, a time-consuming effort in the exam which put additional pressure on them. The extra free-text entry field (as an escape option) was perceived as helpful when no solution could be found in the long menu.

However, students also voiced some concerns about KFPs. For instance, they regretted the fact they had to single out one correct answer phrased in only one or a few words without having the opportunity to elucidate their reflections, as would be the case in an oral exam. The inclusion of a free-text commentary field offered some solution, by allowing students to clarify their choices. Despite these concerns, however, students felt KFPs should remain part of the assessment, as they required knowledge of important and common leading symptoms, which, in turn, sparked intensive self-study. Yet, inclusion of KFPs was conditional on appropriate alignment of assessment contents with those of courses, including VPs.

Statistical analysis

The results of the statistical analysis are presented in Tables 2, 3, and 4. First, the analysis revealed that the two question types did not differ with respect to their level of difficulty (expressed as percentage of correct answers). Compared to crSBAs, the KFP questions presented the highest reliability, as assessed by a G coefficient in a D study. Table 5 gives an overview of the variance components and generalizability coefficients of the KFP and crSBA parts of the 11 exams under investigation. On average, the resolution of one KFP required 169 s, whereas 75 s were needed to solve one crSBA. Considering this, a test based exclusively on KFPs would require 45 min of testing time to reach a 0.80 G coefficient, as opposed to 65 min in a test that would be entirely crSBA-based. The correlation between the two formats was high, implying that both question types measure similar constructs (clinical reasoning). Table 3 presents the reliabilities to be expected for varying test durations, based on hypothetical tests composed of either one of the item types. Table 4 details the correlations of both written exam item types with the total OSCE score. KFPs correlated more closely with the total OSCE score than did crSBAs.

Discussion

In our search for an answer to the question of which assessments formats would be most suitable to test clinical reasoning in clerkship exams, we compared crSBAs and KFPs with respect to students' acceptance, their educational impact and psychometric characteristics in a mixed-methods study. From the analysis of the focus groups four themes emerged reflecting students' perceptions of KFPs compared to their crSBA counterparts: KFPs were perceived as more realistic and difficult, providing a greater stimulus for the intense study of clinical reasoning than did the crSBAs, and were generally well accepted, provided some preconditions were taken into account. The statistical

Table 2. Item- and exam-level characteristics.

Item-level characteristics		
Items	KFP	crSBA
N items (all 11 exams)	81	162
Difficulty: mean (SD)	0.77 (0.13)	0.75 (0.20)
G analysis based on mean variance components over exams		
Items	KFP	crSBA
Mean number of items in the exams (rounded)	7	15
Reliability (G coefficient) for mean number of items in the exams	0.65	0.52
Number of items needed for a G coefficient of 0.80	16	52
Time needed to achieve a 0.80 G coefficient when using one question format only	45 min	65 min
Exam-level characteristics, *N* exams =11		
Correlations	KFP—crSBA	
Correlation: mean (SD)	0.50 (0.11)	
Correlation after Fisher's Z-score transformation	0.50	
True correlation after Fisher's Z-score transformation	0.95	

KFP: Key-feature problem with long menu; crSBA: context-rich single best answer.

Table 3. Reliability per testing time (efficiency); average of 11 exams.

Testing time	1 h	2 h	3 h	4 h
KFP	0.84	0.91	0.94	0.96
crSBA	0.79	0.88	0.92	0.94

Table 4. Correlations.

	OSCE-KFP	OSCE-crSBA
Correlation M (SD)	0.54 (0.12)	0.41 (0.20)
Correlation after Fisher's Z-score transformation M	0.55	0.43
True correlation after Fisher's Z-score transformation M	0.93	0.81

Table 5. Variance components and generalizability coefficients of the key-feature problem (KFP) and context-rich single best answer question (crSBA) parts of the 11 exams under investigation.

		KFP				crSBA			
Exam	N_p[a]	N_i[b]	V_p[c]	V_{pi}[d]	G[e]	N_i[b]	V_p[c]	V_{pi}[d]	G[e]
1	42	8	0.013	0.037	0.73	16	0.004	0.148	0.29
2	34	6	0.005	0.030	0.52	16	0.005	0.144	0.35
3	31	9	0.009	0.039	0.68	14	0.009	0.129	0.48
4	28	6	0.005	0.040	0.45	13	0.019	0.120	0.68
5	31	8	0.006	0.051	0.48	10	0.004	0.161	0.19
6	37	6	0.017	0.037	0.73	10	0.030	0.163	0.65
7	32	7	0.017	0.064	0.65	9	0.010	0.141	0.39
8	33	8	0.012	0.033	0.75	18	0.008	0.094	0.59
9	33	6	0.012	0.042	0.62	17	0.011	0.191	0.48
10	42	8	0.007	0.035	0.61	19	0.014	0.108	0.71
11	33	9	0.010	0.039	0.69	20	0.009	0.157	0.54
Mean[f]		7.4	0.010	0.040	0.65	140.7	0.010	0.140	0.52

[a]Number of exam participants.
[b]Number of items.
[c]Variance component of persons.
[d]Variance component of person x item interaction.
[e]Generalizability coefficient.
[f]Weighted mean (weights: sample size = $N_p \times N_i$).

analysis unveiled no difference in difficulty. KFPs exhibited a higher reliability and efficiency than did the crSBAs. The true correlation of the two written exam parts was high; however, KFPs correlated more closely with the overall OSCE score than did the crSBAs.

Our data indicate that KFPs with a long-menu response format provide a greater stimulus for the intense study of clinical reasoning than do crSBAs. This educational effect of KFPs has not been previously reported for clerkship students. A few studies, however, have demonstrated that the question format, notably the response format, can affect student learning (Frederiksen 1984; Cilliers et al. 2012). Desjardins and colleagues (Desjardins et al. 2014), for instance, recently pointed out that the response format can indeed have an impact on the incidence of cueing, which again can impact perceived difficulty. In their study, they made a comparison between multiple choice-type and open-ended questions, presenting a first group of students with open-ended questions, followed by multiple choice questions, and a second group with the same questions in the reverse order. Irrespective of the format seen first, multiple choice scores resulted higher than those of the open-ended questions. The observed pattern suggests that it was cueing rather than memory for prior questions that led to increased multiple choice questions scores. It may be hypothesized that if students truly believe that crSBAs are easier, in the sense that these require less critical thinking, an examination that is purely crSBA-based could be perceived as less challenging. This could impact the way students prepare for the examination, possibly studying at a more superficial level, which would surely oppose the intent of assessment, which should be to encourage a deep approach to learning (Newble & Jaeger 1983; Al-Kadri et al. 2012). That the response format matters is also supported by a think-aloud protocol study which found that more complex descriptions of thinking patterns are used when solving KFPs with long menu relative to SBAs (Schuwirth et al. 2001).

In terms of psychometric data, we have demonstrated that the measuring capacity of KFPs per unit of testing time is larger than that of crSBAs. With only 7 KFPs we reached a reliability (G coefficient) of 0.65, which means that we would achieve an estimated reliability of 0.80 with only 16 KFPs in a 45-min exam. Hrynchak et al. (2014), in contrast, reported in their review on KFPs aimed at assessing clinical reasoning that internal consistency reliability of KFPs as measured by Cronbach's alpha would be generally acceptable (i.e. between 0.70 and 0.95) if 25–40 KFPs were used (about 3–4 h of testing time). In the case of undergraduate medical students, Fischer et al. (2005) found that a 90-min, 15-KFP exam was able to achieve a reliability of 0.65 (Cronbach's alpha). Similarly, Hatala and Norman (2002) developed a 15-KFP 2-h written exam to assess clinical decision-making skills in internal medicine clerkship students which exhibited an overall test reliability of 0.49. One explanation for the high reliability of the KFPs in the present study could be that KFPs always focused on eight important leading symptoms and were therefore restricted in terms of coverage. Another explanation could be that our KFPs were better aligned with the instruction process (VPs included). Their limited coverage could also explain why, psychometrically, KFPs were of the same difficulty as

crSBAs, where the long-menu format is usually reported to be more difficult (Newble et al. 1979; Veloski et al. 1993).

The fact that KFPs correlated more closely with the total OSCE scores supports the results of the qualitative data. It may be hypothesized that KFPs with a long-menu response format assess active rather than passive knowledge, due to the absence of cueing just like in an OSCE, where there are no predefined answers to choose from either. Further studies are needed, however, to test this assumption. Using VPs for learning (formative assessment) and KFPs for the assessment of clinical reasoning is an example of constructive alignment of goals (to foster active clinical reasoning), methods that support active learning of clinical reasoning (VPs including long-menu questions for self-assessment) and assessment (KFPs which assess clinical reasoning "actively" by long menu) (Biggs 1996).

One of the merits of the present study is that, to our knowledge, it is the first to demonstrate that KFPs with a long-menu response format offer clerkship students a powerful incentive to learn clinical reasoning. Other strengths are the fact that this study was performed in a real setting with many exams and students, and used a mixed-methods design. A limitation of this study is that focus group interviews are generally susceptible to bias, such as that due to the moderator's influence, although we took measures to minimize this. A further limitation is that the coverage of KFPs was restricted to eight leading symptoms and the student population was heterogeneous as it included medical exchange students from abroad, which may have impacted psychometric characteristics. Additionally, the comparability of samples of items over exams was only warranted by the fact that exams were based on the same blueprint and the questions had been generated by the same persons. We would therefore welcome replications of our study *without* such restrictions. Nevertheless, in terms of its educational value, our concept bore fruit in that students put a special emphasis on learning clinical reasoning for the most important leading symptoms as intended. Finally, we encourage studies focusing on one aspect only (either response format or stimulus format or number of questions per case vignette) to enhance our understanding even more.

Conclusions

Students perceived the KFPs with long-menu questions as providing a greater stimulus for the intense study of clinical reasoning than did crSBAs. Statistically, the KFPs revealed a higher efficiency than crSBAs. This study supports the idea that, from an educational perspective, both the stimulus and response format of questions seem to matter. Including KFPs with long menu in clerkship examinations seems to offer valuable opportunities to steer learning in clinical clerkships without psychometric drawbacks.

Acknowledgements

We are grateful to our students who participated in the focus groups. We thank Jörn Heid (Heilbronn University, Heilbronn, Germany) for his support in calculating the times needed to answer questions and Daniel Stricker (Institute of Medical Education, Bern, Switzerland) for providing us with statistical advice.

We thank Angelique van den Heuvel for critically reading and correcting the English manuscript.

Disclosure statement

The authors report no conflict of interest. The authors alone are responsible for the content and writing of this article.

Glossary

Key Feature: A key feature is a critical step in the resolution of a problem, one where examinees are most likely to make a slip when trying to resolve the problem or one that complicates the identification and management of the problem in practice.

(Page GG, Bordage G (1995) The Medical Council of Canada's Key Features Project: a more valid written examination of clinical decision-making skills. Acad Med 1995;70:104–10.)

Key-Feature problem: A key-feature problem is a problem comprising a key feature, which consists of a brief stem with a short patient vignette containing relevant and non-relevant elements, such as symptoms and findings, and is followed by one or more questions.

(Page GG, Bordage G (1995) The Medical Council of Canada's Key Features Project: a more valid written examination of clinical decision-making skills. Acad Med 1995;70:104–10.)

Notes on contributors

Sören Huwendiek, MD, PhD, MME (University of Bern), is a Pediatrician, was one of the Curriculum Coordinators, Chairman of the Centre for Virtual Patients and E-learning Commissioner at Heidelberg Medical Faculty. He is now Head of the Department of Assessment and Evaluation at the Institute of Medical Education in Bern.

Friedrich Reichert, MD, is a Pediatrician at Olgahospital, Klinikum Stuttgart and wrote his MD Thesis about the design of Virtual Patients and their integration into medical curricula.

Cecilia Duncker, MD, is a Pediatric Resident at the University Hospital in Kiel and wrote her MD Thesis about the optimal integration of Virtual Patients into medical curricula.

Bas de Leng, MSc (Medicine) PhD (Medical Education), is an Educational Technologist and Chairman of the E-learning Competency Centre at the Institute of Medical Education, University of Münster, Germany.

Cees PM van der Vleuten, PhD, is a Professor of Education, Scientific Director of the School of Health Professions Education (SHE) at Maastricht University, Maastricht, the Netherlands. He holds honorary appointments in the University of Copenhagen (Denmark), King Saud University (Riyadh) and Radboud University (Nijmegen).

Arno M.M. Muijtjens, MSc, PhD, statistician-methodologist, is an Associate Professor in the Department of Educational Development and Research, Faculty of Health, Medicine and Life Sciences, Maastricht University, the Netherlands.

Hans Martin Bosse, MD, MME, was responsible for alignment with the Pediatric clinical clerkship and is Specialist Registrar in the Department of General Pediatrics, Neonatology and Pediatric Cardiology, University Children's Hospital Düsseldorf, Germany.

Martin Haag, PhD, is a Professor of Software Engineering and Dean at the Faculty of Informatics at Heilbronn University, Germany. Furthermore, he is head of the Centre for Virtual Patients at the University of Heidelberg, Germany and speaker of the working group technology-enhanced learning of the German Society for Medical Informatics.

Georg F. Hoffmann, MD, is a Professor of Pediatrics and Chairman of the University Children's Hospital Heidelberg, Vice Dean of the Medical Faculty of the Ruprecht-Karls-University Heidelberg. He holds honorary

appointments at the Medical Faculties of Padua, Italy, and Tongji, University Wuhan, China.

Burkhard Tönshoff, MD, PhD, is a Professor of Pediatrics and Pediatric Nephrology, holds the position of a Vice Chairman of the Department of Pediatrics I, University Children's Hospital Heidelberg, Germany. He has been involved in the field of Virtual Patients for many years.

Diana H.J.M. Dolmans, PhD, is a Professor of Innovative Learning Arrangements within the School of Health Professions Education/ Department of Educational Development and Research at Maastricht University, the Netherlands.

References

Al-Kadri HM, Al-Moamary MS, Roberts C, van der Vleuten CP. 2012. Exploring assessment factors contributing to students' study strategies: literature review. Med Teach. 34:S42–S50.

Biggs J. 1996. Enhancing teaching through constructive alignment. Higher Educ. 32:347–364. Kluwer Academic Publishers. Printed in the Netherlands.

Braun V, Clarke V. 2006. Using thematic analysis in psychology. Qual Res Psychol. 3:77–101.

Brennan RL. 2001. Generalizability theory. New York (NY): Springer-Verlag.

Bronander KA, Lang VJ, Nixon LJ, Harrell HE, Kovach R, Hingle S, Berman N. 2015. How we developed and piloted an electronic key features examination for the internal medicine clerkship based on a US national curriculum. Med Teach. 37:807–812.

Bordage G, Carretier H, Bertrand R, Page G. 1995. Comparing times and performances of French- and English-speaking candidates taking a national examination of clinical decision-making skills. Acad Med. 70:359–365.

Bosse HM, Dambe R, Juenger J, Kadmon M. 2011. An interdisciplinary and interactive online tool to manage the continuous development of learning objectives in a curriculum. Z Evid Fortbild Qual Gesundhwes. 105:116–123.

Bosse HM, Huwendiek S, Möltner A, Skelin S. (2006): Making a pediatric OSCE fair and reliable. Congress of the Association for Medical Education in Europe (AMEE), Abstract 6O 7, Abstractbooklet, 2006, p. 104.

Cilliers FJ, Schuwirth LW, van der Vleuten CP. 2012. A model of the preassessment learning effects of assessment is operational in an undergraduate clinical context. BMC Med Educ. 12:9.

Desjardins I, Touchie C, Pugh D, Wood TJ, Humphrey-Murto S. 2014. The impact of cueing on written examinations of clinical decision making: a case study. Med Educ. 48:255–261.

Downing SM. 2002. Assessment of knowledge in written test forms. In: Norman G, van der Vleuten C, Newble D, eds. International handbook of research on medical education. Dordrecht, the Netherlands: Kluwer Academic. p. 642–672.

Downing SM. 2009. Written tests: constructed response and selected response formats. In: Downing SM, Yudkowsky R, eds. Assessment in health professions education. New York (NY): Routledge. p. 149–184.

Elstein AS. 1993. Beyond multiple-choice questions and essays: the need for a new way to assess clinical competence. Acad Med. 68:244–249.

Farmer EA, Page G. 2005. A practical guide to assessing clinical decision-making skills using the key features approach. Med Educ. 39:1188–1194.

Fischer MR, Kopp V, Holzer M, Ruderich F, Juenger J. 2005. A modified electronic key feature examination for undergraduate medical students: validation threats and opportunities. Med Teach. 27:450–455.

Frederiksen N. 1984. The real test bias: Influences of testing on teaching and learning. Am Psychol. 39:193–202.

Harden RM, Stevenson M, Downie WW, Wilson GM. 1975. Assessment of clinical competence using objective structured examination. Br Med J. 1:447–451.

Hatala R, Norman GR. 2002. Adapting the key features examination for a clinical clerkship. Med Educ. 36:160–165.

Heemskerk L, Norman G, Chou S, Mintz M, Mandin H, McLaughlin K. 2008. The effect of question format and task difficulty on reasoning strategies and diagnostic performance in internal medicine residents. Adv Health Sci Educ Theory Pract. 13:453–462.

Heid J, Bauch M, Brass K, Hess F, Jünger J, Haag M, Leven FJ. 2006. Entwicklung und Einsatz eines sicheren Prüfungssystems für die medizinische Ausbildung. GMS Med Inform Biom Epidemiol. 2:Doc10.

Higgs J, Jones MA, Loftus S, Christensen N. 2008. Clinical reasoning in the health professions. 3rd ed. London: Elsevier (Butterworth Heinemann).

Huwendiek S, Reichert F, Bosse HM, de Leng BA, van der Vleuten CPM, Haag M, Hoffmann GF, Tönshoff B. 2009. Design principles for virtual patients: a focus group study among students. Med Educ. 43:580–588.

Huwendiek S, Duncker C, Reichert F, de Leng BA, Dolmans D, van der Vleuten CPM, Haag M, Hoffmann GF, Tönshoff B. 2013. Learner preferences regarding integrating, sequencing and aligning virtual patients with other activities in the undergraduate medical curriculum: a focus group study. Med Teach. 35:920–929.

Huwendiek S, Hanebeck B, Bosse HM, Haag M, Hoffmann GF, Tönshoff B. 2007. Lernen und Prüfen mit virtuellen Patienten am Zentrum für Kinder- und Jugendmedizin des Universitätsklinikums Heidelberg: Ergebnisse der Evaluation im Rahmen des E-Learning-Preises Baden-Württemberg 2007. GMS Med Inform Biom Epidemiol. 5:Doc10.

Hrynchak P, Takahashi SG, Nayer M. 2014. Key-feature questions for assessment of clinical reasoning: a literature review. Med Educ. 48:870–883.

Kopp V, Möltner A, Fischer MR. 2006. Key-Feature-Probleme zum Prüfen von prozeduralem Wissen: Ein Praxisleitfaden. GMS Z Med Ausbild. 23:Doc10.

Krueger RA, Casey MA. 2000. Focus groups: a practical guide for applied research. 3rd ed. London: Sage Publications, Inc.

Muchinsky PM. 1996. The correction for attenuation. Educ Psychol Measure. 56:63–75.

McLachlan JC. 2006. The relationship between assessment and learning. Med Educ. 40:716–717.

Newble DI, Jaeger K. 1983. The effect of assessments and examinations on the learning of medical students. Med Educ. 17:165–171.

Newble DI, Baxter A, Elmslie RG. 1979. A comparison of multiple-choice tests and free-response tests in examinations of clinical competence. Med Educ. 13:263–268.

Norman GR, Swanson DB, Case SM. 1996. Conceptual and methodological issues in studies comparing assessment formats. Teach Learn Med. 8:208–216.

Norman G, Bordage G, Page G, Keane D. 2006. How specific is case specificity?. Med Educ. 40:618–623.

Page GG, Bordage G. 1995. The medical council of Canada's key features project: a more valid written examination of clinical decision-making skills. Acad Med. 70:104–110.

Schuwirth LWT, van der Vleuten CPM. 2004. Different written assessment methods: what can be said about their strengths and weaknesses? Med Educ. 38:974–979.

Schuwirth L, van der Vleuten CP, Stoffers HE, Peperkamp AG. 1996. Computerized long-menu questions as an alternative to open-ended questions in computerized assessment. Med Educ. 30:50–55.

Schuwirth LWT, Verheggen MM, van der Vleuten CPM, Boshuizen HPA, Dinant GJ. 2001. Do short cases elicit different thinking processes than factual knowledge questions do? Med Educ. 35:348–356.

van der Vleuten CP, Schuwirth LW. 2005. Assessing professional competence: from methods to programmes. Med Educ. 39:309–317.

Veloski JJ, Rabinowitz HK, Robeson MR. 1993. A solution to the cueing effects of multiple choice questions: the Un-Q format. Med Educ 27:371–375.

Walter KN, Forster J, Scheerer U, Zieger B, Huwendiek S, Bosse HM. 2008. Etablierung eines Prüfungsverbandes Pädiatrie in Baden-Württemberg. Monatsschr Kinderheilkd. 156:100.

Appendix 1: Example of KFP

Vignette:

You are the pediatrician in charge in a pediatric outpatient department. A parent presents the four-year-old girl Jessica who refuses to walk. She has been having recurrent upper respiratory tract infections for four weeks. In the physical exam she appears pale.

Although she refuses to walk, muscular strength, neurologic and joint findings of the lower extremities are normal. All other physical examination is entirely normal.

Question 1:

After having taken a full history and having performed a physical exam:

What would be your next step in the investigation? (Please provide one answer; be as specific as possible.)

Question 2:

In the blood count the following is reported: leucocyte count 2.8/nl (4.5–13/nl); erythrocyte count 2.0/pl (3.9–5.3/pl); hemoglobin 6.1 g/dL (11–14.5 g/dL); hematocrit 18% (31–37%); platelet count 90/nL (180–530/nL). The differential white cell count is still pending.

What is the most likely diagnosis in this case?

Question 3:

How do you confirm your suspected diagnosis?

Example of crSBA

You are a pediatrician in your private practice. A parent presents a three-year-old girl who has suffered from dizziness, nausea and vomiting since two weeks. The physical exam reveals a gait ataxia and postural instability.

What should be done first?

- Physiotherapy
- Referral to a child and adolescent psychiatrist
- Referral to an ophthalmologist and ENT specialist
- Lumbar puncture
- MRI of the brain

Appendix 2: Questioning route

1. How did you perceive the electronic clerkship exam?
2. How did you perceive the different assessment formats in the clerkship exam?
3. Did the different assessment formats have an impact on your learning? If yes, how?