Visualizing cognitive processes in medical education: Forward and backward reasoning in a digital family medicine assessment

Abstract

Background: When evaluating medical students, wouldn't it be helpful to visualize and understand your students' clinical reasoning (CR)? There are different types of CR for students to use in their multiple-choice questions (MCQs). While forward reasoning uses data to generate a hypothesis, backward reasoning looks at possible prompts (answers) to generate a hypothesis. This study implements a new approach to visualizing CR during digital MCQ assessment. Furthermore, it examines the effect of feedback given during the learning process, also known as formative feedback, on the thought process.

Methods: Quantitative and qualitative data were collected at the end of two consecutive year 5 end-of-semester Family Medicine exams in 2023. Both exams consisted of 60 MCQs and an additional research component also comprising MCQs. During the research component students digitally recorded their reasoning process while answering MCQs. Qualitative data were coded via three rounds of coding, including two marking/coding parties.

Results: This study was able to digitally visualize CR in a large cohort (n=210). On average, the exam questions were answered with 87% of CR. Forward reasoning was used significantly more often than backward reasoning (WS 22/23 p=0.006, SS 23 p<0.001). High performers were significantly more likely to use forward reasoning and backward reasoning than low performers (p<0.01). Formative Feedback had no significant influence on the choice of CR type (p=0.281). Follow-up questions might stimulate a change in CR behaviour; however further research is needed (p<0.001).

Conclusion: This study illustrates an alternative method to visualize students' cognitive processes on a large scale. This approach sheds light on the required cognitive processes. It may help educators to better understand what to focus on during curricular learning activities aimed at preparing for state exams. This method may be beneficial as a quality criterion for MCQ questions, as it does not only rely on expert opinion or question metrics but illustrates students' cognitive processes when answering MCQ.

Keywords: clinical reasoning, family medicine, forward reasoning, backward reasoning, undergraduate medical education

Johanna Klutmann[1]
Constanze Dietzsch[1]
Ute Schlasius-Ratter[2]
Alexander Oksche[2]
Sara Volz-Willems[1]
Sandra Jordan[1]
Johannes Jäger[1]
Fabian Dupont[1]

1 Saarland University, Department of Family Medicine, Homburg (Saar), Germany

2 German Institute for State Examinations in Medicine, Pharmacy, Dentistry and Psychotherapy (IMPP), Mainz, Germany

Introduction

Clinical reasoning (CR) is a core competency in medical education representing the thought process behind diagnosing and treating patients [1], [2], [3], [4], [5]. However, visualizing or assessing CR, especially in undergraduate medical education, remains a challenge [6].

Although CR is central to medical practice and education, traditional assessment methods - especially MCQs, which dominate high-stakes exams, are criticized for not properly capturing the cognitive processes involved in CR [7], [8]. Despite innovations like Key-feature Questions (KFQs), little is known about how CR can actually be mapped or visualized within standard MCQ assessments [9].

At the same time, many national licensing exams wish to maintain MCQs for their objectivity, standardization, and cost efficiency [10], [11]. Given the widespread use and high stakes MCQ assessment it's crucial to understand and visualize how these questions stimulate or reflect CR during those assessment periods [12], [13]. Mapping CR during these assessments may significantly enhance both educational quality and clinical preparedness for students transitioning into the clinical field.

Previous research has distinguished between forward reasoning (FR) and backward reasoning (BR) [14] with FR being often described as the hallmark of expertise and deeper understanding [15]. FR describes the thought process in which students answer questions without having to read through the choices, generating their hypothesis from the MCQ question stem and possible added material [3]. BR describes thinking backwards and drawing on the answer choices (distractors) to answer the question instead [3]. The distinction between FR and BR only captures one aspect of how CR can be described. It focuses on how the reasoning process was formed. However, CR is a multidimensional construct. It can also be understood in terms of its goals, performance and contextual factors, all of which can be the focus of analysis [2], [4].

Approaches such as KFQs and Formative Feedback (FF) have been introduced to promote CR during assessments [9]. KFQs focus on one difficult aspect of solving a problem, often embedding this feature in a written case followed by a limited number of questions [16]. FF is a core element of "assessment for learning" [17]. It provides feedback during the learning process with a focus on assisting learning [17]. FF deepens students' understanding, even during assessments [18].

To date, educators use MCQ metrics to describe MCQ items and their quality. These metrics offer valuable insights into how items perform on a population level – identifying which questions are too easy, too hard, or particularly effective at distinguishing between high- and low-performing students [19]. However, these measures are independent of students' cognitive thought processes [20]. They provide no information about how or why a student arrived at a particular answer. Exploring observable reasoning processes rather than MCQ metrics could provide new insights into how CR is elicited, assessed, and supported during MCQ-based exams.

Other studies have called for a better understanding of CR utilization during assessment [21], [22]. This study investigates whether CR can be visualized during an undergraduate MCQ exam and how it relates to performance. Additionally, the influence of FF and follow up key feature questions (KFQ) on the usage of CR are analysed. The aim of this study is to examine whether CR processes can be identified in the context of an undergraduate MCQ examination and to explore their relationship to student performance. More specifically, we analyse whether the use of CR, in particular FR and BR, is associated with higher performance across different item types and if and to what extent FF and follow-up key feature questions (KFQ) elicit different reasoning strategies.

Methods

Setting and study participants

In this mixed methods study, all participants were year five undergraduate medical students at Saarland University (UdS). The exam was the end of year compulsory Family Medicine exam, a digital state exam based MCQ assessment. It contained two identical exam setups including 60 MCQs, at the end of the winter semester 2022/2023 (WS 22/23) and at the end of the summer semester 2023 (SS 23). Both exams contained a research component, which consisted of two KFQs. These KFQs were case studies with follow-up questions. The KFQs used were not the same. Firstly, to include a wider range of questions in the study. Secondly, to prevent students from already knowing the questions due to possible student-internal discussions about the questions. In this study, the questions will be referred to by an abbreviation. The first number indicates the case study number and the second indicates the follow-up question number. After each follow-up question of the KFQ, there was an open-ended text box question, asking for a self-assessment of whether the students derived the question clinically and their cognitive thought process used in the previous question. The same structure applied to the SS23 exam. Additionally, in SS23, students received FF for the first time during an exam for a selection of questions. The FF was uniform information-based feedback consisting of the correct answer and a brief explanation. The participants were randomly divided into two groups: A (n=63) and B (n=52). Group A received FF after each follow-up question of the second KFQ and Group B received FF after each follow-up question of the first KFQ. In both groups, FF was provided in reverse order, to maintain face validity and comparability of CR (with and without FF). Ethical clearance was provided before the study (234/20-14.04.2022). The participants consented to the use of their study and exam performance before the exam.

MCQ item selection

In cooperation with the Institute for State Examinations in Medicine, Pharmacy, Dentistry and Psychotherapy (IMPP), a panel of two research students (JK, CD) and four Family Medicine faculty (SJ, SVW, FD, JJ) selected the follow-up KFQs from an MCQ pool provided by the IMPP based on the learning objectives of the semester. The follow-up KFQs used were divided into procedural, diagnostic, and factual questions.

Data collection and analysis

Both quantitative and qualitative data were collected and then exported to Excel (version 16.96.1). Qualitative data from both exams were analysed using three levels of deductive content analysis (see figure 1), informed by a structured literature review. First, coders applied Young et al.'s framework to determine the presence of CR [2]. The aforementioned study identified six categories of terms regarding CR on which there was a consensus. These were: the purpose/goal of reasoning; the outcome of reasoning; reasoning performance; reasoning processes; reasoning skills; and the context of reasoning [2].

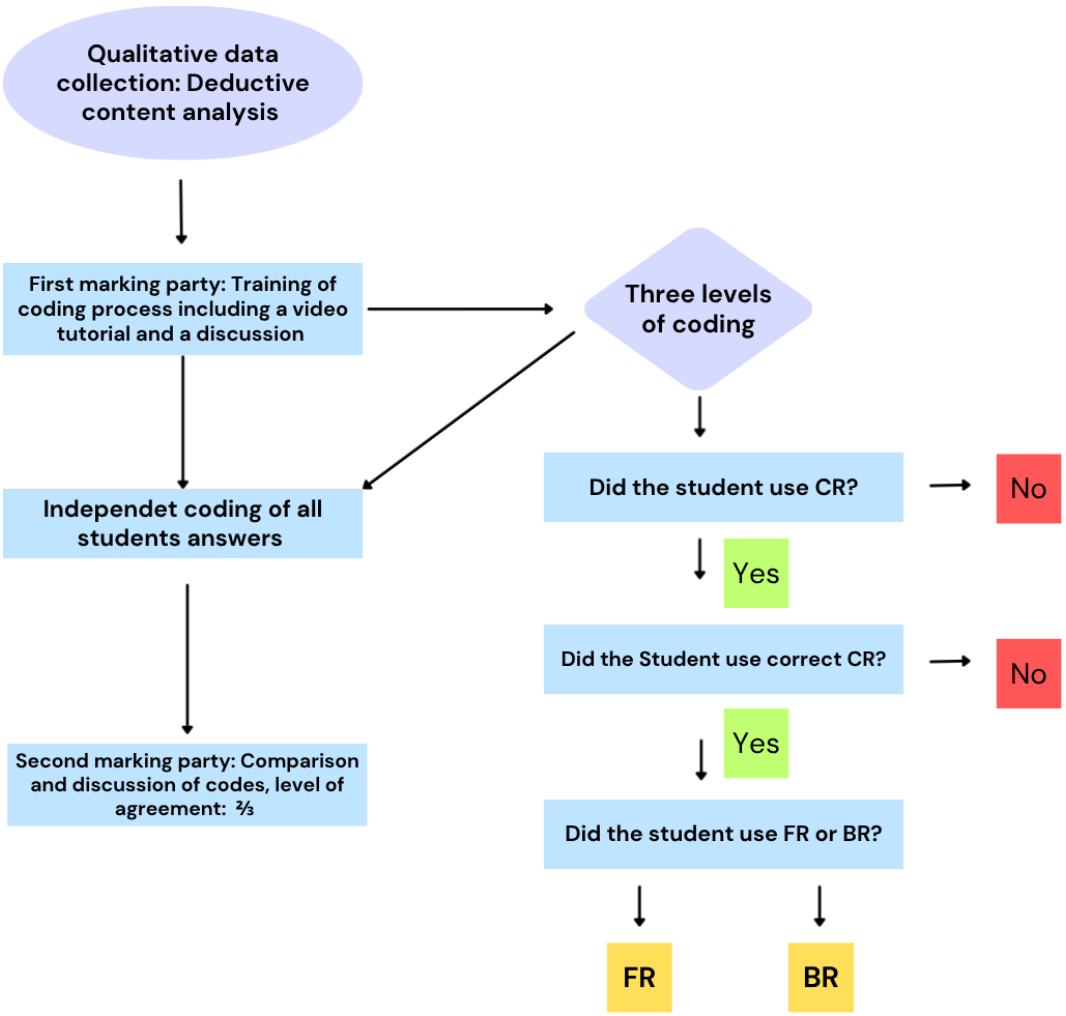

Figure 1: Qualitative deductive content analysis with three levels of coding

These categories and their associated themes served as a guideline to determine CR in our study. Second, they assessed the correctness and coherence of reasoning, focusing on clarity and conclusiveness rather than accuracy alone. Correct CR did not only mention an action, but the action is justified in the context of clinical logic. Whereas in an incorrect CR response essential clinical considerations were missing or overlooked and therefore a wrong conclusion was made. Third, correct CRs were categorized as FR or BR following Beullens et al. [3]. This study gave an example to distinguish FR and BR: An example of a data-driven reasoning [BR] statement is: "If he has an elevated blood sugar, then he must have diabetes". An example of a hypothesis-driven reasoning [FR] statement is: "Because he has diabetes, he has an elevated blood sugar" [3]. The coding guidelines also included the aspect to analyse if students justify their answers solely based on the answers provided (BR), or if the answer resulted from a more extensive clinical thought process (FR). The application of preexisting factual knowledge was coded as FR as a fast form of CR. This was based on the aspect of the Croskerry model, that repeated system 2 analyses, can become automatic system 1 responses (e.g., pattern recognition) [23]. Initially, coders (JK, CD, USR, SH, FD, JJ) were trained via video and discussion. All had prior qualitative coding experience. Each response was coded independently by all six coders. In a second session, codes with at least 4/6 agreement were accepted; discrepancies were discussed until full consensus was reached.

Quantitative data analysis was performed with jamovi™ version 2.4.12. The analysis included descriptive statistics of FR and BR frequencies, a median split to define performance groups, and (χ^2) tests to examine group differences and CR types regarding follow-up questions and FF. Binomial/multinomial logistic regressions tested whether follow-up questions or FF predicted CR type.

Results

The following results describe patterns of CR usage and differences in performance across two cohorts of students taking exams. It also examines the effects of feedback on one of these cohorts. In WS 22/23 95 out of 97 participated, in SS 23 115 out of 115 participated. The average performance of the exam was comparable for the two semesters, with average scores of 80% (WS 22/23) and 84% (SS 23) respectively. In total 4.95% of the an-

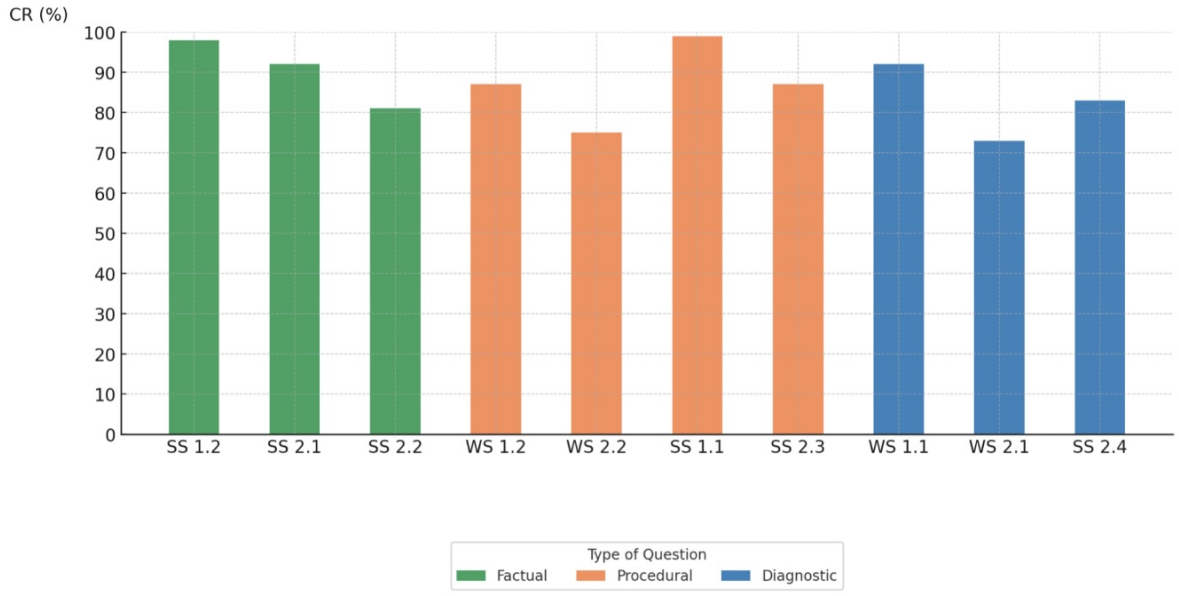

Figure 2: CR usage by question type and exam. Exam questions are categorized by semester (WS or SS) and number of the KFQ case study and follow-up question

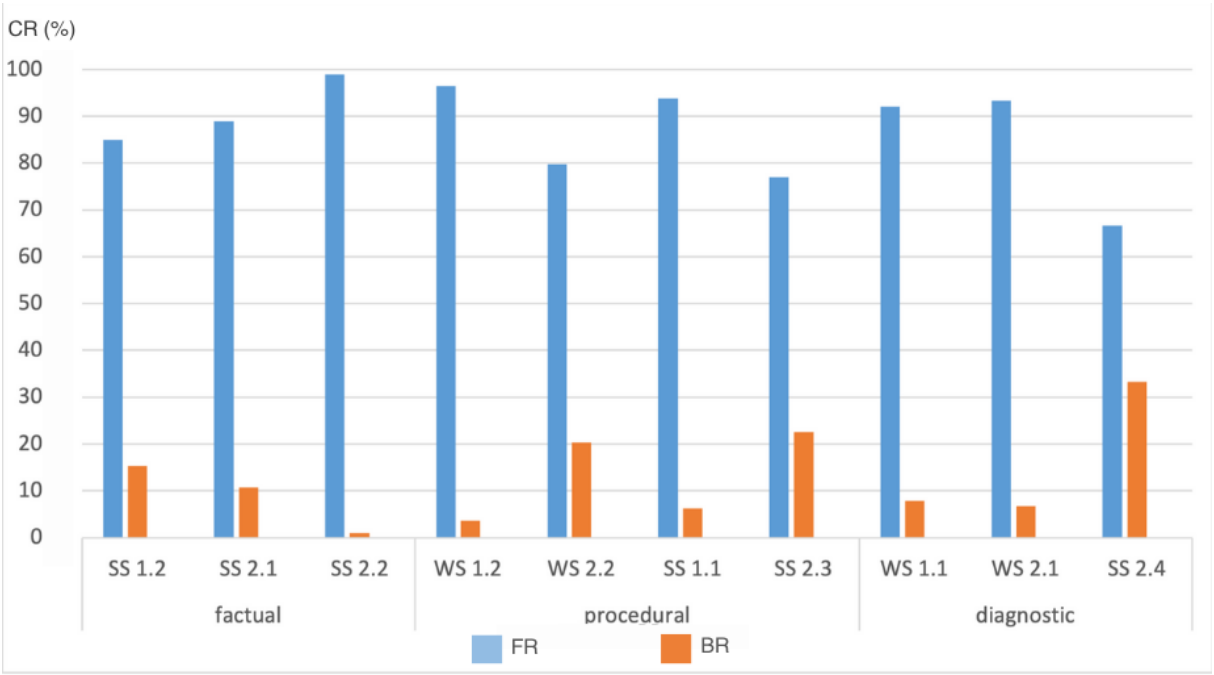

Figure 3: FR (Forward Reasoning) and BR (Backward Reasoning) usage compared between the three question types (factual, procedural, diagnostic) in summer- and winter semester (WS or SS) and number of the KFQ case study and follow-up question

swers in WS 22/23 and 1.44% of the answers in SS23 were skipped.

The use of clinical reasoning

CR was observed across both exams. An example of an answer coded as CR regarding the diagnosis of an obstructive lung disease, was: "I looked at the Tiffeneau Index. Furthermore, inhalation of salbutamol did not result in any reversibility of the obstruction. This led me to the diagnosis of obstructive ventilation disorder." (WS.21.40). An exemplary response coded as No CR regarding the same question was "[I] cannot clearly identify either obstruction or restriction; therefore [I] simply took a mixture." (WS.21.73).

On average, the exam questions were answered with 87% CR. As can be seen in figure 2, the highest proportion of CR usage was found in a procedural question (99.1%), while the lowest was noted in a diagnostic question (73.5%). In all questions, CR was applied more frequently than no CR.

The highest proportion of FR usage was found in a factual question (98.9%), while the lowest was noted in a diagnostic question (66.7%) (see figure 3). For BR, the

highest usage was in a diagnostic question (33.3%) and the lowest in a factual question (1.1%). FR occurred significantly more often than BR, both in WS 22/23 (p=0.006) and SS 23 (p<0.001). Two examples of coding for FR and BR regarding the further course of action in acute appendicitis were for FR: "Appendicitis should be treated surgically as quickly as possible; therefore, fasting upon admission to hospital" (SS.11.39) and for BR: "None of the other possible answers would have been appropriate in the case of acute appendicitis" (SS.11.96).
The fast form of FR, which consisted of the application of preexisting knowledge was coded in 3.5% of the FR answers.

High performers were more likely to apply CR

Students scoring at or above the median were classified as high performers and those below as low performers. High performers were more likely to apply CR (FR and BR) than low performers. The odds of using FR and BR were 3.3 (OR=3.33, 95% CI [2.378, 4.490], p<0.001) and 2.7 times higher (OR=2.67, 95% CI [1.669, 4.261], p<0.001), respectively, for high performers. In one question in WS 22/23, the high performers exclusively used FR. Moreover, students who applied any type of CR were more likely to answer a question correctly compared to students who did not use CR (p<0.001).

Feedback did not influence the use of CR

A binomial logistic regression showed no significant correlation between FF and the use of CR (χ^2=1.78, p=0.182). CR occurred neither more frequently after FF nor after no FF. A multinomial logistic regression showed no significant influence of FF on the choice of CR type (χ^2=3.38, p=0.281). Neither the comparison between BR and "no CR" (p=0.099), nor the comparison between FR and "no CR" (p=0.227) reached statistical significance.

Higher probability of BR in follow-up KFQs in one semester

In SS 23, a significant difference was found in the frequency of FR and BR between the first questions (1.1, 2.1) and the follow-up KFQ (1.2, 2.2, 2.3, 2.4) (χ^2=256, p< 0.001). Students applied BR significantly more often for follow-up questions than for the first question. The probability of BR was about 2.5 times higher when it was a follow-up question compared to a first question (OR=2.49, 95% CI [2.29, 2.84], p<0.001). In WS 22/23, there was no significant difference in the frequency of FR and BR in relation to the first or follow-up questions (χ^2=1.08, p=0.299).

Discussion

This study yielded four main findings. First, family medicine students were able to demonstrate CR within a tablet-based assessment format. Second, high-performing students were more likely to apply CR, and the use of CR was associated with a better performance. Third, FF had not significantly influenced the type of CR employed. Fourth, follow-up questions were associated with an increased likelihood of BR.
The results demonstrate that CR can be effectively captured and analysed in a digital assessment environment. CR processes were visualized across an entire academic year, enabling simultaneous data collection from a large cohort during the examination. This approach offers a valuable alternative to traditional post-exam interviews and may reveal more efficiently students' CR processes. Contrary to common assumptions, the data suggests that medical students do not simply focus on answer options in MCQs [24]. Instead, many actively engage in reasoning processes, as FR was used significantly more frequently than BR. In a factual question, FR was used most often. This may be attributed to the fact that a large proportion of students remember factual answers from learnt information. The application of preexisting factual knowledge was coded as FR as a fast form of CR. This is comparable to pattern recognition of the Croskerry model [23]. Fewer students used FR in more complex questions. Consistent with existing literature, high performers relied more on FR than low performers [3]. This supports the notion that FR is more typical for expert reasoning patterns [15]. This could mean that the use of FR is situation dependent and that its application may increase with advancing competency levels (experts). These findings reinforce the importance of CR not only in clinical practice but also in undergraduate assessments. When students employed CR to answer a question, the likelihood of a correct answer increased. There is also evidence that FR may be more common among students who are able to think flexibly and apply knowledge creatively. For example, in WS 22/23, in one question high performers exclusively used FR. According to the panel that selected the questions, this item required "thinking outside the box" and went beyond standard diagnostic or therapeutic content, promoting students to draw on clinical imagination or real-world experience such as internships.
However, modern question types such as KFQs, which include follow-up questions, do not inherently promote FR. On the contrary, follow-up questions can lead students to prematurely focus on a single hypothesis and the preset answer options. This phenomenon, known as premature closure, limits the consideration of differential diagnoses and leads to search satisfaction – where reasoning halts after identifying one plausible diagnosis [25]. This may explain the increased likelihood of BR in response to follow-up questions in SS23. On the other hand, in WS 22/23, there was no significant difference in the frequency of FR and BR between the first and follow-up questions. This revealed a difference between the two

cohorts, illustrating how the structure and type of assessment items can influence the reasoning process. The use of different questions, coupled with an imbalanced mix of factual, diagnostic and procedural questions highlights the importance of carefully considering item design in medical examinations. Seemingly minor variations in question format can substantially influence whether students engage in FR or BR strategies. Even though both semesters were subject to the same family medicine curriculum, it is important to consider the possibility that the reasoning strategies employed in the two semesters may have been influenced by differing practical experiences. In addition, disparities in learning motivation or stress management may underpin the observed variation in the frequency of BR. In situations of uncertainty, the propensity to avoid errors may be augmented, thereby facilitating a process of thinking backwards or the promotion of BR. Group psychology that develops within a cohort through exchange, mentors, or so-called "exam myths" could also influence the reasoning strategies of the cohorts. While innovative competency-based formats may aim to assess higher-order reasoning, they do not automatically elicit authentic CR.

Similarly, the use of FF did not enhance CR in this study. While FF is known to support learning [17], [18], the design of the FF may influence its effectiveness. In the assessment, students received uniform information-based feedback consisting of the correct answer and a brief explanation. Research shows that CR skills improve when feedback includes specific suggestions for improvement, regardless of students' performance [26]. The lack of individualization or absence of suggestions for improvement may explain why the FF did not significantly influence the usage of CR or FR.

To confirm these findings, further research with larger and more diverse cohorts at different stages of undergraduate medical education is needed. Future studies should explore the impact of KFQ formats and varied feedback types across disciplines, while continuing to visualize CR at scale.

Limitations

The study focused on year five medical students within a Family Medicine curriculum in Germany. Family medicine, being a broad and integrative subject, may have contributed to the high point average (average scores of 80% and 84%). This could explain the high use of CR in general and FR in particular. Consequently, the data of this study may not be generalizable to students in other years, disciplines, or institutions. Furthermore, reasoning types (FR and BR) were coded based on students' written explanation of their reasoning, focusing on their initial train of thought. Although six independent coders were involved and inter-rater agreement of at least two-thirds was achieved for all responses, it is possible that nuances of reasoning were lost in the process. There is also potential for desirability bias. Another limitation of this study is that factual questions were only present in one of the two cohorts (SS23). The examination questions were selected from an IMPP item pool according to curricular relevance, while the categorization into factual, diagnostic, and procedural domains was only applied retrospectively for the analysis. Future studies should include a more balanced set of questions to enable advanced comparisons between cohorts.

Conclusion

The findings indicate that careful assessment setup choices can and should be used to foster conscious CR choices in students. Digital assessments have additional potential to effectively visualize CR. This large-scale CR visualization may complement the quality screening of MCQ, especially for future high-stakes exams. Further research into large scale CR visualization may help to better understand student performance differences and it may help individualize standardized assessments. CR visualization may also help align MCQ CR with clinical practice and therefore increase its value for workplace training.

Abbreviations

- WS: Winter semester
- SS: Summer semester
- CR: Clinical reasoning
- FR: Forward reasoning
- BR: Backward reasoning
- IMPP: Institut für medizinische und pharmazeutische Prüfungsfragen (German Institute for State Examinations in Medicine, Pharmacy, Dentistry and Psychotherapy (IMPP))
- MCQ: Multiple choice question
- KFQ: Key feature question
- Wonca: World Organisation of National Colleges of Family Medicine

Authors' ORCIDs

- Alexander Oksche: [0000-0003-4592-1770]
- Fabian Dupont: [0000-0003-2247-5640]

Competing interests

The authors declare that they have no competing interests.

FD is current Wonca World executive member and director & Wonca World young doctor lead.

References

1. Hawks MK, Maciuba JM, Merkebu J, Durning SJ, Mallory R, Arnold MJ, Torre D, Soh M. Clinical Reasoning Curricula in Preclinical Undergraduate Medical Education: A Scoping Review. Acad Med. 2023;98(8):958-965. DOI: 10.1097/ACM.0000000000005197
2. Young M, Thomas A, Gordon D, Gruppen L, Lubarsky S, Rencic J, Ballard T, Holmboe E, Da Silva A, Ratcliffe T, Schuwirth L, Durning SJ. The terminology of clinical reasoning in health professions education: implications and considerations. Med Teach. 2019;41(11):1277-1284. DOI: 10.1080/0142159X.2019.1635686
3. Beullens J, Struyf E, Van Damme B. Do extended matching multiple-choice questions measure clinical reasoning? Med Educ. 2005;39(4):410-417. DOI: 10.1111/j.1365-2929.2005.02089.x
4. Connor DM, Durning SJ, Rencic JJ. Clinical reasoning as a core competency. Acad Med. 2020;95(8):1166-1171. DOI: 10.1097/ACM.0000000000003027
5. Eva KW. What every teacher needs to know about clinical reasoning. Med Educ. 2005;39(1):98-106. DOI: 10.1111/j.1365-2929.2004.01972.x
6. Guth TA, Wolfe RM, Martinez O, Subhiyah RG, Henderek JJ, McAllister C, Roussel D. Assessment of Clinical Reasoning in Undergraduate Medical Education: A Pragmatic Approach to Programmatic Assessment. Acad Med. 2024;99(8):912-921. DOI: 10.1097/ACM.0000000000005665
7. Daniel M, Rencic J, Durning SJ, Holmboe E, Santen SA, Lang V, Ratcliffe T, Gordon D, Heist B, Lubarsky S, Estrada CA, Ballard T, Artino Jr AR, Sergio Da Silva A, Cleary T, Stojan J, Gruppen LD. Clinical Reasoning Assessment Methods: A Scoping Review and Practical Guidance. Acad Med. 2019;94(6):902-912. DOI: 10.1097/ACM.0000000000002618
8. Mee J, Pandian R, Wolczynski J, Morales A, Paniagua M, Harik P, Baldwin P, Clauser BE. An experimental comparison of multiple-choice and short-answer questions on a high-stakes test for medical students. Adv Health Sci Educ Theory Pract. 2024;29(3):783-801. DOI: 10.1007/s10459-023-10266-3
9. Hrynchak P, Glover Takahashi S, Nayer M. Key-feature questions for assessment of clinical reasoning: A literature review. Med Educ. 2014;48(9):870-883. DOI: 10.1111/medu.12509
10. Ricketts C, Brice J, Coombes L. Are multiple choice tests fair to medical students with specific learning disabilities? Adv Health Sci Educ Theory Pract. 2010;15(2):265-275. DOI: 10.1007/s10459-009-9197-8
11. Chenot JF. Undergraduate medical education in Germany. Ger Med Sci. 2009;7:Doc02. DOI: 10.3205/000061
12. Salam A, Yousuf R, Bakar SA. Multiple choice questions in medical education: how to construct high quality questions. Int J Hum Health Sci. 2020;4(2):79. DOI: 10.31344/ijhhs.v4i2.180
13. Law AK, So J, Lui CT, Choi YF, Cheung KH, Hung KK, Graham CA. AI versus human-generated multiple-choice questions for medical education: a cohort study in a high-stakes examination. BMC Med Educ. 2025;25(1):208. DOI: 10.1186/s12909-025-06796-6
14. Patel VL, Groen GJ, Arocha JF. Medical expertise as a function of task difficulty. Mem Cognit. 1990;18(4):394-406. DOI: 10.3758/bf03197128
15. Shin HS. Reasoning processes in clinical reasoning: from the perspective of cognitive psychology. Korean J Med Educ. 2019;31(4):299-308. DOI: 10.3946/kjme.2019.140
16. Al-Wardy NM. Assessment methods in undergraduate medical education. Sultan Qaboos Univ Med J. 2010;10(2):203-209.
17. Murugesan M, David PL, Chitra CB. Correlation between Formative and Summative Assessment Results by Post Validation in Medical Undergraduates. IOSR J Dent Med Sci. 2021;20(9):51-57. DOI: 10.9790/0853-2009055157
18. Badyal DK, Bala S, Singh T, Gulrez G. Impact of immediate feedback on the learning of medical students in pharmacology. J Adv Med Educ Prof. 2019;7(1):1-6. DOI: 10.30476/JAMP.2019.41036
19. Chauhan GR, Chauhan BR, Vaza JV, Chauhan PR. Relations of the Number of Functioning Distractors With the Item Difficulty Index and the Item Discrimination Power in the Multiple Choice Questions. Cureus. 2023;15(7):e42492. DOI: 10.7759/cureus.42492
20. Rao C, Kishan Prasad H, Sajitha K, Permi H, Shetty J. Item analysis of multiple choice questions: Assessing an assessment tool in medical students. Int J Educ Psychol Res. 2016;2(4):201-204. DOI: 10.18203/2320-6012.ijrms20161256
21. Heist BS, Gonzalo JD, Durning S, Torre D, Elnicki DM. Exploring Clinical Reasoning Strategies and Test-Taking Behaviors During Clinical Vignette Style Multiple-Choice Examinations: A Mixed Methods Study. J Grad Med Educ. 2014;6(4):709-714. DOI: 10.4300/JGME-D-14-00176.1
22. Torre D, Daniel M, Ratcliffe T, Durning SJ, Holmboe E, Schuwirth L. Programmatic Assessment of Clinical Reasoning: New Opportunities to Meet an Ongoing Challenge. Teach Learn Med. 2025;37(3):403-411. DOI: 10.1080/10401334.2024.2333921
23. Croskerry P. Critical thinking and reasoning in emergency medicine. In: Croskerry P, Cosby KS, Schenkel SM, Wears RL, editors. Patient safety in emergency medicine. Philadelphia, PA: Lippincott Williams & Wilkins; 2008. p.213-218.
24. van Wijk EV, Janse RJ, Ruijter BN, Rohling JH, van der Kraan J, Crobach S, de Jonge M, de Beaufort AJ, Dekker FW, Lagers AM. Use of very short answer questions compared to multiple choice questions in undergraduate medical students: an external validation study. PLoS One. 2023;18(7):e0288558. DOI: 10.1371/journal.pone.0288558
25. Prakash S, Bihari S, Need P, Sprick C, Schuwirth L. Immersive high fidelity simulation of critically ill patients to study cognitive errors: a pilot study. BMC Med Educ. 2017;17(1):36. DOI: 10.1186/s12909-017-0871-x
26. De Beer M, Mårtensson L. Feedback on students' clinical reasoning skills during fieldwork education. Aust Occup Ther J. 2015;62(4):255-264. DOI: 10.1111/1440-1630.12208

Corresponding author:
Prof. Dr. med. Fabian Dupont, MHPE
Saarland University, Department of Family Medicine, Geb. 80.2, D-66424 Homburg (Saar), Germany
Fabian.dupont@uks.eu

Please cite as
Klutmann J, Dietzsch C, Schlasius-Ratter U, Oksche A, Volz-Willems S, Jordan S, Jäger J, Dupont F. Visualizing cognitive processes in medical education: Forward and backward reasoning in a digital family medicine assessment. GMS J Med Educ. 2026;43(5):Doc65.
DOI: 10.3205/zma001859, URN: urn:nbn:de:0183-zma0018594

This article is freely available from
https://doi.org/10.3205/zma001859

Received: *2025-07-15*
Revised: *2025-11-21*
Accepted: *2026-01-08*
Published: *2026-06-15*

Visualisierung kognitiver Prozesse in der medizinischen Ausbildung: Vorwärts- und Rückwärtsdenken in einer digitalen Prüfung der Allgemeinmedizin

Johanna Klutmann[1]
Constanze Dietzsch[1]
Ute Schlasius-Ratter[2]
Alexander Oksche[2]
Sara Volz-Willems[1]
Sandra Jordan[1]
Johannes Jäger[1]
Fabian Dupont[1]

1 Universität des Saarlandes, Zentrum Allgemeinmedizin, Homburg (Saar), Deutschland

2 Institut für medizinische und pharmazeutische Prüfungsfragen (IMPP), Mainz, Deutschland

Zusammenfassung

Hintergrund: Wäre es bei der Bewertung von Medizinstudierenden nicht hilfreich, das klinische Denken (CR) Ihrer Studierenden zu visualisieren und zu verstehen? Es gibt verschiedene Arten von CR, die Studierende in ihren Multiple-Choice-Fragen (MCQs) verwenden können. Während beim Vorwärtsdenken Daten zur Erstellung einer Hypothese verwendet werden, werden beim Rückwärtsdenken mögliche Hinweise (Antworten) zur Erstellung einer Hypothese herangezogen. Diese Studie implementiert einen neuen Ansatz zur Visualisierung von CR während der digitalen MCQ-Prüfung. Darüber hinaus untersucht sie die Auswirkungen von Feedback während des Lernprozesses, auch bekannt als formatives Feedback, auf den Denkprozess.

Methoden: Quantitative und qualitative Daten von zwei Semesterabschlussprüfungen im Fach Allgemeinmedizin wurden am Ende des 5. Studienjahres im Jahr 2023 gesammelt. Beide Prüfungen bestanden aus 60 MCQs und einem zusätzlichen Forschungsteil, welcher ebenfalls MCQs umfasste. Während des Forschungsteils erfassten die Studierenden ihren Denkprozess bei der Beantwortung der MCQs digital. Die qualitativen Daten wurden in drei Kodierungsrunden, darunter zwei Markierungs-/Kodierungspartys, kodiert.

Ergebnisse: Diese Studie konnte CR in einer großen Kohorte (n=210) digital visualisieren. Im Durchschnitt wurden die Prüfungsfragen mit 87% CR beantwortet. Vorwärtsdenken wurde signifikant häufiger verwendet als Rückwärtsdenken (WS 22/23 $p=0,006$, SS 23 $p<0,001$). Leistungsstarke Studierende verwendeten signifikant häufiger Vorwärtsdenken und Rückwärtsdenken als leistungsschwache Studierende ($p<0,01$). Formatives Feedback hatte keinen signifikanten Einfluss auf die Wahl der CR-Art ($p=0,281$). Folgefragen könnten eine Veränderung des CR-Verhaltens bewirken; jedoch sind weitere Untersuchungen erforderlich ($p<0,001$).

Schlussfolgerung: Diese Studie veranschaulicht eine alternative Methode zur Visualisierung der kognitiven Prozesse von Studierenden in großem Maßstab. Dieser Ansatz beleuchtet die erforderlichen kognitiven Prozesse. Er kann Pädagogen dabei helfen, besser zu verstehen, worauf sie sich bei curricularen Lernaktivitäten zur Vorbereitung auf staatliche Prüfungen konzentrieren sollten. Diese Methode kann als Qualitätskriterium für MCQ-Fragen von Vorteil sein, da sie sich nicht nur auf Expertenmeinungen oder Fragenmetriken stützt, sondern auch die kognitiven Prozesse der Studierenden bei der Beantwortung von MCQs veranschaulicht.

Schlüsselwörter: klinisches Denken, Allgemeinmedizin, Vorwärtsdenken, Rückwärtsdenken, medizinische Ausbildung im Grundstudium

Einleitung

Klinisches Denken (Clinical Reasoning, CR) ist eine Kernkompetenz in der medizinischen Ausbildung, die den Denkprozess hinter der Diagnosefindung und Behandlung von Patienten und Patientinnen darstellt [1], [2], [3], [4], [5]. Die Visualisierung und Bewertung von CR, insbesondere in der medizinischen Grundausbildung, bleibt jedoch eine Herausforderung [6].

Obwohl CR für die medizinische Praxis und Ausbildung von zentraler Bedeutung ist, werden traditionelle Bewertungsmethoden – insbesondere MCQs, die bei wichtigen Prüfungen dominieren – dafür kritisiert, dass sie die kognitiven Prozesse, die bei CR eine Rolle spielen, nicht angemessen erfassen [7], [8]. Trotz Innovationen wie Key-Feature-Fragen (KFQs) ist wenig darüber bekannt, wie CR tatsächlich innerhalb von Standard-MCQ-Bewertungen abgebildet oder visualisiert werden kann [9].

Gleichzeitig möchten viele nationale Zulassungsprüfungen MCQs aufgrund ihrer Objektivität, Standardisierung und Kosteneffizienz beibehalten [10], [11]. Angesichts der weit verbreiteten Verwendung und der hohen Bedeutung von MCQ-Prüfungen ist es von entscheidender Bedeutung, zu verstehen und zu visualisieren, wie diese Fragen das CR während dieser Bewertungsphasen stimulieren oder widerspiegeln [12], [13]. Die Abbildung von CR während Prüfungen kann die Qualität der Ausbildung als auch die klinische Vorbereitung von Studierenden für den klinischen Bereich erheblich verbessern.

Frühere Forschungen haben zwischen Vorwärtsdenken (Forward Reasoning, FR) und Rückwärtsdenken (Backward Reasoning, BR) unterschieden [14], wobei FR oft als Kennzeichen für Fachwissen und tieferes Verständnis beschrieben wird [15]. FR beschreibt den Denkprozess, bei dem Studierende Fragen beantworten, ohne die Antwortmöglichkeiten durchlesen zu müssen, und ihre Hypothese aus der MCQ-Frage und möglichen zusätzlichen Informationen ableiten [3]. BR beschreibt das Rückwärtsdenken und das Zurückgreifen auf die Antwortmöglichkeiten (Distraktoren), um die Frage zu beantworten [3]. Die Unterscheidung zwischen FR und BR erfasst nur einen Aspekt der Beschreibung von CR. Sie konzentriert sich darauf, wie der Denkprozess entstanden ist. CR ist jedoch ein mehrdimensionales Konstrukt. Es kann auch im Hinblick auf seine Ziele, seine Leistung und kontextuellen Faktoren verstanden werden, die alle im Mittelpunkt der Analyse stehen können [2], [4].

Ansätze wie KFQs und Formatives Feedback (FF) wurden eingeführt, um CR während der Bewertung zu fördern [9]. KFQs konzentrieren sich auf einen schwierigen Aspekt der Problemlösung und betten dieses Merkmal häufig in einen schriftlichen Fall ein, gefolgt von einer begrenzten Anzahl von Fragen [16]. FF ist ein Kernelement der „Bewertung für das Lernen" („assessment for learning") [17]. Es liefert Feedback während des Lernprozesses mit dem Schwerpunkt auf der Unterstützung des Lernens [17]. FF vertieft das Verständnis der Studierenden, sogar während der Prüfungen [18].

Bislang verwenden Pädagogen und Pädagoginnen MCQ-Metriken, um MCQ-Items und deren Qualität zu beschreiben. Diese Metriken bieten wertvolle Einblicke in die Leistung von Items auf Populationsebene – sie identifizieren, welche Fragen zu einfach, zu schwer oder besonders effektiv sind, um zwischen leistungsstarken und leistungsschwachen Studierenden zu unterscheiden [19]. Diese Messungen sind jedoch unabhängig von den kognitiven Denkprozessen der Studierenden [20]. Sie liefern keine Informationen darüber, wie oder warum Studierende zu einer bestimmten Antwort gelangt sind. Die Untersuchung beobachtbarer Denkprozesse könnte neue Erkenntnisse darüber liefern, wie CR während MCQ-basierter Prüfungen hervorgerufen, bewertet und unterstützt wird.

Andere Studien fordern ein besseres Verständnis der Nutzung von CR während der Prüfung [21], [22].

Das Ziel dieser Studie ist es, zu untersuchen, ob CR-Prozesse im Kontext einer MCQ-Prüfung für Studierende identifiziert werden können, und den Zusammenhang dieser mit der Leistung der Studierenden zu erforschen. Genauer gesagt analysieren wir, ob die Verwendung von CR, insbesondere FR und BR, mit einer höheren Leistung bei verschiedenen Aufgabentypen verbunden ist und ob und inwieweit FF und Folgefragen der KFQs unterschiedliche Denkstrategien hervorrufen.

Methodik

Setting und Studienteilnehmende

In dieser Mixed-Methods-Studie waren alle Teilnehmer und Teilnehmerinnen Studierende im fünften Jahr des Medizinstudiums an der Universität des Saarlandes (UdS). Die Prüfung war die obligatorische Jahresabschlussprüfung im Fach Allgemeinmedizin, eine digitale MCQ-Prüfung mit IMPP-Fragen (Institut für medizinische und pharmazeutische Prüfungsfragen). Sie umfasste zwei identische Prüfungssettings mit jeweils 60 MCQs, die am Ende des Wintersemesters 2022/2023 (WS 22/23) und am Ende des Sommersemesters 2023 (SS 23) durchgeführt wurden. Beide Prüfungen enthielten einen Forschungsteil, der aus zwei KFQs bestand. Bei diesen KFQs handelte es sich um kurze Fallbeschreibungen mit Folgefragen. Die verwendeten KFQs waren nicht identisch. Erstens, um ein breiteres Spektrum an Fragen in die Studie einzubeziehen. Zweitens, um zu verhindern, dass die Studierenden die Fragen durch Diskussionen zwischen den Semestern bereits kannten. In dieser Studie werden die Fragen mit einer Abkürzung bezeichnet. Die erste Zahl gibt die Nummer der Fallstudie an, die zweite die Nummer der Folgefrage. Nach jeder Folgefrage des KFQ folgte eine offene Textfeldfrage, in der die Studierenden selbst einschätzen sollten, ob sie die Frage klinisch hergeleitet hatten und welchen kognitiven Denkprozess sie bei der vorherigen Frage angewendet hatten. Die gleiche Struktur galt für die SS23-Prüfung. Zusätzlich erhielten die Studierenden in SS23 zum ersten Mal während einer Prüfung

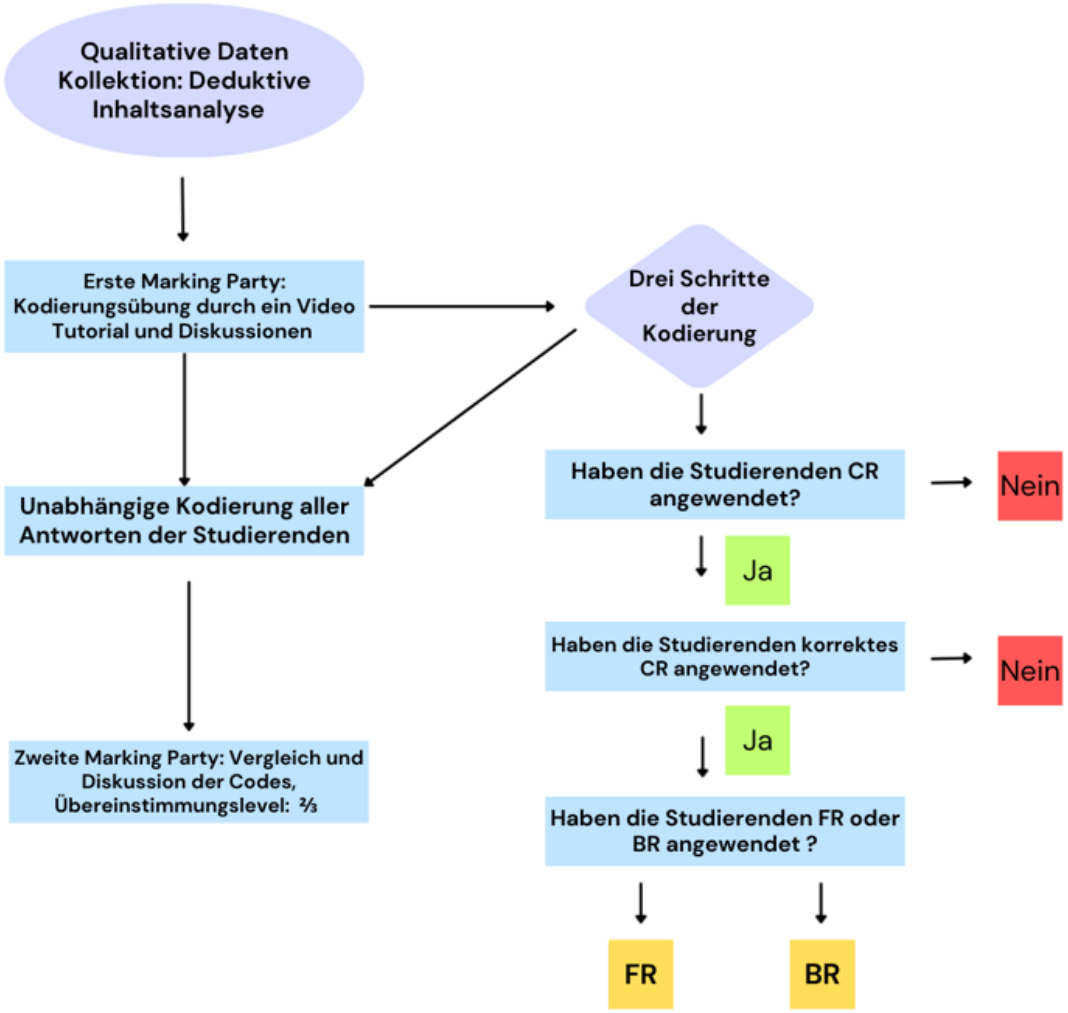

Abbildung 1: Qualitative deduktive Inhaltsanalyse mit drei Kodierungsebenen

für eine Auswahl von Fragen FF. Das FF war ein einheitliches, informationsbasiertes Feedback, das aus der richtigen Antwort und einer kurzen Erklärung bestand. Die Teilnehmenden wurden zufällig in zwei Gruppen eingeteilt: A (n=63) und B (n=52). Gruppe A erhielt FF nach jeder Folgefrage des zweiten KFQ und Gruppe B erhielt FF nach jeder Folgefrage des ersten KFQ. In beiden Gruppen wurde FF in umgekehrter Reihenfolge bereitgestellt, um die Augenscheinvalidität und Vergleichbarkeit von CR (mit und ohne FF) zu gewährleisten. Die ethische Unbedenklichkeit wurde vor der Studie bestätigt (234/20-14.04.2022). Die Teilnehmenden stimmten vor der Prüfung der Verwendung ihrer Studien- und Prüfungsleistungen zu.

Auswahl der MCQ-Fragen

In Zusammenarbeit mit dem IMPP wählte ein Gremium aus zwei Forschungsstudierenden (JK, CD) und vier Fakultätsmitgliedern der Allgemeinmedizin (SJ, SVW, FD, JJ) die Folge-KFQs aus einem vom IMPP bereitgestellten MCQ-Pool auf der Grundlage der Lernziele des Semesters aus. Die verwendeten Folge-KFQs wurden in Prozedere-, Diagnose- und Sachfragen unterteilt.

Datenerhebung und -analyse

Es wurden sowohl quantitative als auch qualitative Daten erhoben und anschließend in Excel (Version 16.96.1) exportiert. Die qualitativen Daten aus beiden Prüfungen wurden anhand einer strukturierten Literaturrecherche mit Hilfe einer dreistufigen deduktiven Inhaltsanalyse (siehe Abbildung 1) ausgewertet. Zunächst wandten die Kodierer das Rahmenkonzept von Young et al. an, um das Vorhandensein von CR zu bestimmen [2]. Die genannte Studie identifizierte sechs Kategorien von Begriffen in Bezug auf CR, über die Konsens bestand. Diese waren: Zweck/Ziel des Denkens; Ergebnis des Denkens; Denkleistung; Denkprozesse; Denkfähigkeiten; und Kontext des Denkens [2]. Diese Kategorien und die damit verbundenen Themen dienten als Leitfaden zur Bestimmung von CR in unserer Studie. Zweitens bewerteten sie die Richtigkeit und Kohärenz des Denkens. Hierbei lag der Fokus auf Klarheit und Schlüssigkeit der Argumentation als auf alleiniger Genauigkeit. Korrekte CR erwähnten nicht nur eine Handlung, sondern die Handlung wurde auch im Kontext der klinischen Logik begründet. In einer falschen CR-Antwort fehlten hingegen wesentliche klinische Überlegungen oder wurden übersehen, sodass eine falsche Schlussfolgerung gezogen wurde. Drittens wurden

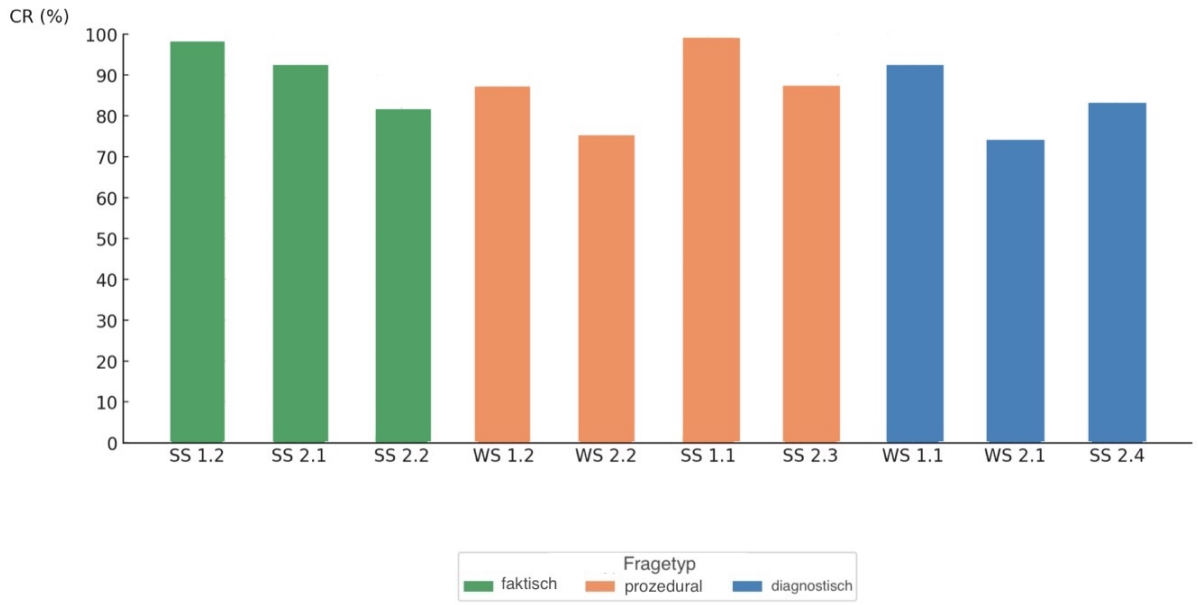

Abbildung 2: Verwendung von CR nach Fragetyp und Prüfung. Die Prüfungsfragen sind nach Semester (WS oder SS) sowie nach Nummer der KFQ-Fallstudie und der Folgefrage kategorisiert

korrekte CRs nach Beullens et al. [3] als FR oder BR kategorisiert. Diese Studie gab ein Beispiel zur Unterscheidung von FR und BR: „Ein Beispiel für eine datengestützte Argumentation [BR] ist: „Wenn er einen erhöhten Blutzucker hat, dann muss er Diabetes haben". Ein Beispiel für eine hypothesengetriebene Argumentation [FR] ist: „Weil er Diabetes hat, hat er einen erhöhten Blutzucker" [3]. Die Kodierungsrichtlinien umfassten auch den Aspekt, zu analysieren, ob die Studierenden ihre Antworten ausschließlich auf der Grundlage der vorgegebenen Antworten (BR) begründen oder ob die Antwort das Ergebnis eines umfassenderen klinischen Denkprozesses (FR) ist. Die Anwendung bereits vorhandenen Faktenwissens wurde als FR als schnelle Form von CR kodiert. Dies basierte auf dem Aspekt des Croskerry-Modells, dass wiederholte System-2-Analysen zu automatischen System-1-Reaktionen werden können (z. B. Mustererkennung) [23]. Zunächst wurden die Kodierer (JK, CD, USR, SH, FD, JJ) per Video und Diskussion geschult. Alle hatten bereits Erfahrung mit qualitativer Kodierung. Jede Antwort wurde von allen sechs Kodierern unabhängig voneinander kodiert. In einer zweiten Sitzung wurden Codes mit mindestens 4/6 Übereinstimmungen akzeptiert; Diskrepanzen wurden diskutiert, bis ein vollständiger Konsens erzielt wurde.

Die quantitative Datenanalyse wurde mit jamovi™ Version 2.4.12 durchgeführt. Die Analyse umfasste deskriptive Statistiken der FR- und BR-Häufigkeiten, eine Medianaufteilung der Punktzahl zur Definition von Leistungsgruppen und (χ^2)-Tests zur Untersuchung von Gruppenunterschieden und CR-Typen in Bezug auf Folgefragen und FF. Binomiale/multinomiale logistische Regressionen testeten, ob Folgefragen oder FF CR-Typen vorhersagen können.

Ergebnisse

Die folgenden Ergebnisse beschreiben Muster der CR-Nutzung und Leistungsunterschiede zwischen zwei Kohorten von Studierenden, die Prüfungen abgelegt haben. Außerdem werden die Auswirkungen von Feedback auf eine dieser Kohorten untersucht. Im WS 22/23 nahmen 95 von 97 Studierenden teil, im SS 23 nahmen 115 von 115 Studierenden teil. Die durchschnittliche Leistung der Prüfung war in beiden Semestern vergleichbar, mit Durchschnittswerten von 80% (WS 22/23) bzw. 84% (SS 23). Insgesamt wurden 4,95% der Fragen im WS 22/23 und 1,44% der Fragen im SS 23 nicht beantwortet.

Die Anwendung von klinischem Denken

CR wurde in beiden Prüfungen beobachtet. Ein Beispiel für eine als CR kodierte Antwort bezüglich der Diagnose einer obstruktiven Lungenerkrankung war: „Ich [habe] mir den Tiffeneau Index angeschaut [...]. Außerdem hat die Inhalation von Salbutamol zu keiner Reversibilität der Obstruktion ergeben. Somit kam ich auf den Befund der obstruktiven Ventilationsstörung." (WS.21.40). Eine beispielhafte Antwort, die in Bezug auf dieselbe Frage als „kein CR" kodiert wurde, lautete: „Ich finde weder Obstruktion noch Restriktion eindeutig daher [habe ich] einfach [eine] Mischung genommen." (WS.21.73).

Im Durchschnitt wurden die Prüfungsfragen mit 87% CR beantwortet. Wie aus Abbildung 2 hervorgeht, war der höchste Anteil an CR-Verwendungen bei einer Prozedere-frage zu verzeichnen (99,1%), während der niedrigste Anteil bei einer Diagnosefrage festgestellt wurde (73,5%). Bei allen Fragen wurde CR häufiger angewendet als kein CR.

Der höchste Anteil an FR-Verwendung wurde bei einer Sachfrage (98,9%) festgestellt, während der niedrigste

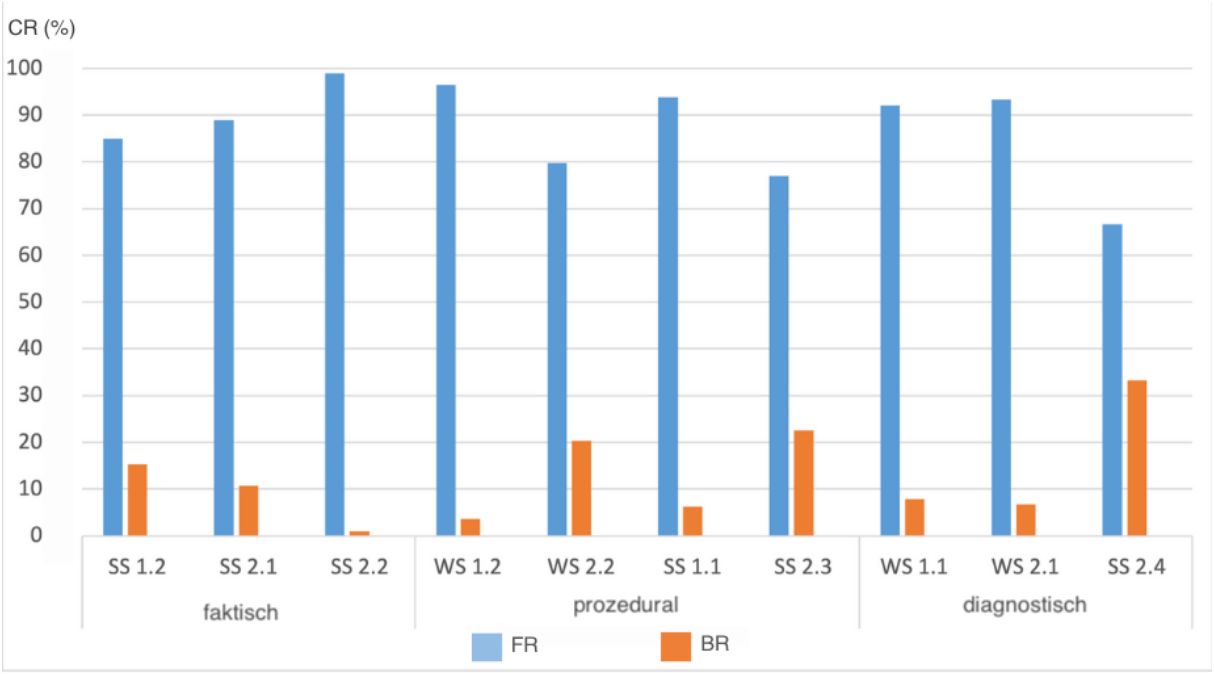

Abbildung 3: Vergleich der Anwendung von FR (Vorwärtsschlussfolgerung) und BR (Rückwärtsschlussfolgerung) bei den drei Fragetypen (sachbezogene, prozedurale und diagnostische Fragen) im Sommer- und Wintersemester (WS bzw. SS) sowie Nummer der KFQ-Fallstudie und der Folgefrage

bei einer diagnostischen Frage (66,7%) zu verzeichnen war (siehe Abbildung 3). Bei BR war die höchste Verwendung bei einer diagnostischen Frage (33,3%) und die niedrigste bei einer Sachfrage (1,1%) zu verzeichnen. FR kam sowohl in WS 22/23 (p=0,006) als auch in SS 23 (p<0,001) signifikant häufiger vor als BR. Zwei Beispiele für die Kodierung von FR und BR hinsichtlich der weiteren Vorgehensweise bei einer akuten Blinddarmentzündung waren für FR: „Eine Appendizitis sollte schnellstmöglich operativ therapiert werden, daher eine Nahrungskarenz bei [Krankenhaus] Einweisung" (SS.11.39) und für BR: „Alle anderen angegeben Antwortmöglichkeiten wären bei einer akuten Appendizitis nicht angemessen gewesen" (SS.11.96).
Die schnelle Form von FR, die aus der Anwendung bereits vorhandenen Wissens bestand, wurde in 3,5% der FR-Antworten kodiert.

Leistungsstarke wandten CR häufiger an

Studierende, deren Punktzahl dem Median entsprach oder diesen überstieg, wurden als Leistungsstarke eingestuft, diejenigen darunter als Leistungsschwache. Leistungsstarke wandten CR (FR und BR) häufiger an als Leistungsschwache. Die Wahrscheinlichkeit, FR und BR zu verwenden, betrug 3,3 (OR=3,33, 95% KI [2,378, 4,490], p<0,001) bzw. war 2,7-mal höher (OR=2,67, 95% KI [1,669, 4,261], p<0,001) für Leistungsstarke. Bei einer Frage in WS 22/23 verwendeten die Leistungsstarken ausschließlich FR. Darüber hinaus beantworteten Studierende, die irgendeine Art von CR anwendeten, eine Frage mit höherer Wahrscheinlichkeit richtig als Studierende, die keine CR verwendeten (p<0,001).

Feedback hatte keinen Einfluss auf die Verwendung von CR

Eine binomiale logistische Regression zeigte keine signifikante Korrelation zwischen FF und der Verwendung von CR (χ^2=1,78, p=0,182). CR kam weder nach FF noch nach keinem FF häufiger vor. Eine multinomiale logistische Regression zeigte keinen signifikanten Einfluss von FF auf die Wahl der CR-Art (χ^2=3,38, p=0,281). Weder der Vergleich zwischen BR und „kein CR" (p=0,099) noch der Vergleich zwischen FR und „kein CR" (p=0,227) erreichten statistische Signifikanz.

Höhere Wahrscheinlichkeit von BR in Folge-KFQs in einem Semester

In SS 23 wurde ein signifikanter Unterschied in der Häufigkeit von FR und BR zwischen den ersten Fragen (1.1, 2.1) und den Folge-KFQs (1.2, 2.2, 2.3, 2.4) festgestellt (χ^2=256, p< 0,001). Die Studierenden wandten BR bei Folgefragen signifikant häufiger an als bei der ersten Frage. Die Wahrscheinlichkeit von BR war bei Folgefragen etwa 2,5-mal höher als bei der ersten Frage (OR=2,49, 95% KI [2,29, 2,84], p<0,001). Im WS 22/23 gab es keinen signifikanten Unterschied in der Häufigkeit von FR und BR in Bezug auf die ersten oder Folgefragen (χ^2=1,08, p=0,299).

Diskussion

Diese Studie ergab vier wesentliche Erkenntnisse. Erstens konnten Studierende der Allgemeinmedizin CR innerhalb eines Tablet-basierten Bewertungsformats demonstrieren. Zweitens wandten leistungsstarke Studierende CR eher an, und die Anwendung von CR war mit einer besseren Leistung verbunden. Drittens hatte FF keinen signifikanten Einfluss auf die Art der angewandten CR. Viertens waren Folgefragen mit einer erhöhten Wahrscheinlichkeit von BR verbunden.

Die Ergebnisse zeigen, dass CR in einer digitalen Bewertung effektiv erfasst und analysiert werden kann. CR-Prozesse wurden über eine gesamte Semesterpopulation hinweg visualisiert, was die gleichzeitige Datenerfassung einer großen Kohorte während der Prüfung ermöglichte. Dieser Ansatz bietet eine wertvolle Alternative zu traditionellen Interviews nach der Prüfung und kann die CR-Prozesse der Studierenden effizienter aufzeigen. Entgegen der allgemeinen Annahme deuten die Daten darauf hin, dass Medizinstudierende sich nicht einfach auf die Antwortmöglichkeiten in MCQs konzentrieren [24]. Stattdessen beschäftigen sich viele aktiv mit Denkprozessen, da FR deutlich häufiger als BR verwendet wurde. Bei einer Sachfrage wurde FR am häufigsten verwendet. Dies kann darauf zurückgeführt werden, dass sich ein großer Teil der Studierenden an sachliche Antworten aus gelernten Informationen erinnert. Die Anwendung von bereits vorhandenem Sachwissen wurde als FR als schnelle Form des CR kodiert. Dies ist vergleichbar mit der Mustererkennung des Croskerry-Modells [23]. Bei komplexeren Fragen verwendeten weniger Studierende FR. In Übereinstimmung mit der vorhandenen Literatur verließen sich leistungsstarke Studierende stärker auf FR als leistungsschwache Studierende [3]. Dies stützt die Annahme, dass FR eher für die Argumentationsmuster von Experten typisch ist [15]. Dies könnte bedeuten, dass die Verwendung von FR situationsabhängig ist und dass seine Anwendung mit steigendem Kompetenzniveau (Experten) zunimmt. Diese Ergebnisse unterstreichen die Bedeutung von CR nicht nur in der klinischen Praxis, sondern auch bei der Bewertung von Studierenden. Wenn Studierende CR zur Beantwortung einer Frage einsetzten, stieg die Wahrscheinlichkeit einer richtigen Antwort. Die Studie ergab Hinweise darauf, dass FR bei Studierenden, die flexibel denken und Wissen kreativ anwenden können, häufiger vorkommt. Im WS 22/23 verwendeten beispielsweise leistungsstarke Studierende bei einer Frage ausschließlich FR. Nach Ansicht des Gremiums, das die Fragen ausgewählt hatte, erforderte diese Frage „unkonventionelles Denken" und ging über den üblichen diagnostischen oder therapeutischen Inhalt hinaus, sodass die Studierenden ihre klinische Vorstellungskraft oder praktische Erfahrungen, beispielsweise aus Praktika, einbringen mussten.

Moderne Fragetypen wie KFQs, die Folgefragen enthalten, fördern jedoch nicht von Natur aus FR. Im Gegenteil, Folgefragen können dazu führen, dass sich Studierende vorzeitig auf eine einzige Hypothese und die vorgegebenen Antwortoptionen konzentrieren. Dieses als „vorzeitiger Abschluss" bekannte Phänomen schränkt die Berücksichtigung von Differentialdiagnosen ein und führt zu „Suchbefriedigung" – wobei das Denken nach der Identifizierung einer plausiblen Diagnose zum Stillstand kommt [25]. Dies könnte die erhöhte Wahrscheinlichkeit von BR als Antwort auf Folgefragen in SS23 erklären. Andererseits gab es im WS 22/23 keinen signifikanten Unterschied in der Häufigkeit von FR und BR zwischen den ersten und den Folgefragen. Dies zeigt einen Unterschied zwischen den beiden Kohorten und veranschaulichte, wie die Struktur und die Art der Prüfungsaufgaben den Denkprozess beeinflussen können. Die Verwendung unterschiedlicher Fragen in Verbindung mit einer unausgewogenen Mischung aus sachlichen, diagnostischen und prozeduralen Fragen unterstreicht, wie wichtig es ist, die Gestaltung der Fragen in medizinischen Prüfungen sorgfältig zu überdenken. Scheinbar geringfügige Abweichungen im Frageformat können einen erheblichen Einfluss darauf haben, ob Studierende FR- oder BR-Strategien anwenden. Auch wenn beide Semester dem gleichen Lehrplan für Allgemeinmedizin unterlagen, muss berücksichtigt werden, dass die in den beiden Semestern angewandten Argumentationsstrategien möglicherweise durch unterschiedliche praktische Erfahrungen beeinflusst wurden. Darüber hinaus könnten Unterschiede in der Lernmotivation oder im Stressmanagement die beobachteten Abweichungen in der Häufigkeit von BR begründen. In Situationen der Unsicherheit kann die Neigung, Fehler zu vermeiden, verstärkt sein, was einen Prozess des Rückwärtsdenkens oder die Förderung von BR begünstigt. Die Gruppenpsychologie, die sich innerhalb einer Kohorte durch Austausch, Mentoren oder sogenannte „Prüfungsmythen" entwickelt, könnte ebenfalls die Denkstrategien der Kohorten beeinflussen. Innovative kompetenzbasierte Formate zielen zwar darauf ab, das Denken auf höherer Ebene zu bewerten, führen jedoch nicht automatisch zu authentischem CR.

Ebenso hat der Einsatz von FF in dieser Studie das CR nicht verbessert. Obwohl FF bekanntermaßen das Lernen unterstützt [17], [18], kann die Gestaltung des FF dessen Wirksamkeit beeinflussen. Bei der Bewertung erhielten die Studierenden einheitliches, informationsbasiertes Feedback, das aus der richtigen Antwort und einer kurzen Erklärung bestand. Untersuchungen zeigen, dass sich die CR-Fähigkeiten verbessern, wenn das Feedback spezifische Verbesserungsvorschläge enthält, unabhängig von der Leistung der Studierenden [26]. Der Mangel an Individualisierung oder das Fehlen von Verbesserungsvorschlägen könnte erklären, warum das FF keinen signifikanten Einfluss auf die Nutzung von CR oder FR hatte.

Um diese Ergebnisse zu bestätigen, sind weitere Untersuchungen mit größeren und vielfältigeren Kohorten in verschiedenen Phasen des Medizinstudiums erforderlich. Zukünftige Studien sollten die Auswirkungen von KFQ-Formaten und unterschiedlichen Arten von Feedback in verschiedenen Fachbereichen untersuchen und die Visualisierung von CR in großem Maßstab weiterführen.

Limitationen

Die Studie konzentrierte sich auf Medizinstudierende im fünften Jahr im Fach Allgemeinmedizin in Deutschland. Die Allgemeinmedizin ist ein breites und integratives Fachgebiet, was möglicherweise zu den hohen durchschnittlichen Leistungen (Durchschnittswerten von 80% und 84%) beigetragen hat. Dies könnte die häufige Verwendung von CR im Allgemeinen und FR im Besonderen erklären. Folglich lassen sich die Daten dieser Studie möglicherweise nicht auf Studierende anderer Jahrgänge, Fachrichtungen oder Institutionen übertragen. Darüber hinaus wurden die Argumentationstypen (FR und BR) auf der Grundlage der schriftlichen Erläuterungen der Studierenden zu ihrer Argumentation kodiert, wobei der Schwerpunkt auf ihrem ursprünglichen Gedankengang lag. Obwohl sechs unabhängige Kodierer und Kodiererinnen beteiligt waren und für alle Antworten eine Übereinstimmung von mindestens zwei Dritteln erreicht wurde, ist es möglich, dass dabei Nuancen der Argumentation verloren gegangen sind. Es besteht auch die Möglichkeit einer Verzerrung aufgrund von sozialer Erwünschtheit (desirability bias). Eine weitere Einschränkung dieser Studie besteht darin, dass Faktenfragen nur in einer der beiden Kohorten (SS23) vorkamen. Die Prüfungsfragen wurden anhand ihrer Relevanz für den Lehrplan aus einem IMPP-Fragenpool ausgewählt, während die Einteilung in Fakten-, Diagnose- und Verfahrensbereiche erst nachträglich für die Analyse vorgenommen wurde. Zukünftige Studien sollten einen ausgewogeneren Fragenkatalog enthalten, um weitergehende Vergleiche zwischen den Kohorten zu ermöglichen.

Schlussfolgerung

Die Ergebnisse deuten darauf hin, dass sorgfältige Entscheidungen bei der Gestaltung der Prüfungen getroffen werden können und sollten, um bewusste CR-Entscheidungen bei den Studierenden zu fördern. Digitale Prüfungsformate bieten zusätzliches Potenzial für eine effektive Visualisierung des CR. Diese groß angelegte CR-Visualisierung kann die Qualitätsprüfung von MCQ ergänzen, insbesondere bei zukünftigen Prüfungen mit hoher Bedeutung. Weitere Forschungen zur groß angelegten CR-Visualisierung können zu einem besseren Verständnis der Leistungsunterschiede zwischen den Studierenden und zur Individualisierung standardisierter Bewertungen beitragen. Die CR-Visualisierung kann auch dazu beitragen, die MCQs an die klinische Praxis anzupassen und somit ihren Wert für die Ausbildung am Arbeitsplatz zu steigern.

Abkürzungen

- WS: Wintersemester
- SS: Sommersemester
- CR: klinisches Denken/ Clinical reasoning
- FR: Vorwärtsdenken/ Forward reasoning
- BR: Rückwärtsdenken/ Backward reasoning
- IMPP: Institut für medizinische und pharmazeutische Prüfungsfragen
- MCQ: Multiple Choice Question
- KFQ: Key Feature Question
- Wonca: World Organisation of National Colleges of Family Medicine

ORCIDs der Autoren

- Alexander Oksche: [0000-0003-4592-1770]
- Fabian Dupont: [0000-0003-2247-5640]

Interessenkonflikt

Die Autor*innen erklären, dass sie keinen Interessenkonflikt im Zusammenhang mit diesem Artikel haben.
FD ist derzeit Vorstandsmitglied von Wonca World sowie Direktor und Leiter des Bereichs „Junge Ärzte“ bei Wonca World.

Literatur

1. Hawks MK, Maciuba JM, Merkebu J, Durning SJ, Mallory R, Arnold MJ, Torre D, Soh M. Clinical Reasoning Curricula in Preclinical Undergraduate Medical Education: A Scoping Review. Acad Med. 2023;98(8):958-965. DOI: 10.1097/ACM.0000000000005197
2. Young M, Thomas A, Gordon D, Gruppen L, Lubarsky S, Rencic J, Ballard T, Holmboe E, Da Silva A, Ratcliffe T, Schuwirth L, Durning SJ. The terminology of clinical reasoning in health professions education: implications and considerations. Med Teach. 2019;41(11):1277-1284. DOI: 10.1080/0142159X.2019.1635686
3. Beullens J, Struyf E, Van Damme B. Do extended matching multiple-choice questions measure clinical reasoning? Med Educ. 2005;39(4):410-417. DOI: 10.1111/j.1365-2929.2005.02089.x
4. Connor DM, Durning SJ, Rencic JJ. Clinical reasoning as a core competency. Acad Med. 2020;95(8):1166-1171. DOI: 10.1097/ACM.0000000000003027
5. Eva KW. What every teacher needs to know about clinical reasoning. Med Educ. 2005;39(1):98-106. DOI: 10.1111/j.1365-2929.2004.01972.x
6. Guth TA, Wolfe RM, Martinez O, Subhiyah RG, Henderek JJ, McAllister C, Roussel D. Assessment of Clinical Reasoning in Undergraduate Medical Education: A Pragmatic Approach to Programmatic Assessment. Acad Med. 2024;99(8):912-921. DOI: 10.1097/ACM.0000000000005665
7. Daniel M, Rencic J, Durning SJ, Holmboe E, Santen SA, Lang V, Ratcliffe T, Gordon D, Heist B, Lubarsky S, Estrada CA, Ballard T, Artino Jr AR, Sergio Da Silva A, Cleary T, Stojan J, Gruppen LD. Clinical Reasoning Assessment Methods: A Scoping Review and Practical Guidance. Acad Med. 2019;94(6):902-912. DOI: 10.1097/ACM.0000000000002618
8. Mee J, Pandian R, Wolczynski J, Morales A, Paniagua M, Harik P, Baldwin P, Clauser BE. An experimental comparison of multiple-choice and short-answer questions on a high-stakes test for medical students. Adv Health Sci Educ Theory Pract. 2024;29(3):783-801. DOI: 10.1007/s10459-023-10266-3

9. Hrynchak P, Glover Takahashi S, Nayer M. Key-feature questions for assessment of clinical reasoning: A literature review. Med Educ. 2014;48(9):870-883. DOI: 10.1111/medu.12509

10. Ricketts C, Brice J, Coombes L. Are multiple choice tests fair to medical students with specific learning disabilities? Adv Health Sci Educ Theory Pract. 2010;15(2):265-275. DOI: 10.1007/s10459-009-9197-8

11. Chenot JF. Undergraduate medical education in Germany. Ger Med Sci. 2009;7:Doc02. DOI: 10.3205/000061

12. Salam A, Yousuf R, Bakar SA. Multiple choice questions in medical education: how to construct high quality questions. Int J Hum Health Sci. 2020;4(2):79. DOI: 10.31344/ijhhs.v4i2.180

13. Law AK, So J, Lui CT, Choi YF, Cheung KH, Hung KK, Graham CA. AI versus human-generated multiple-choice questions for medical education: a cohort study in a high-stakes examination. BMC Med Educ. 2025;25(1):208. DOI: 10.1186/s12909-025-06796-6

14. Patel VL, Groen GJ, Arocha JF. Medical expertise as a function of task difficulty. Mem Cognit. 1990;18(4):394-406. DOI: 10.3758/bf03197128

15. Shin HS. Reasoning processes in clinical reasoning: from the perspective of cognitive psychology. Korean J Med Educ. 2019;31(4):299-308. DOI: 10.3946/kjme.2019.140

16. Al-Wardy NM. Assessment methods in undergraduate medical education. Sultan Qaboos Univ Med J. 2010;10(2):203-209.

17. Murugesan M, David PL, Chitra CB. Correlation between Formative and Summative Assessment Results by Post Validation in Medical Undergraduates. IOSR J Dent Med Sci. 2021;20(9):51-57. DOI: 10.9790/0853-2009055157

18. Badyal DK, Bala S, Singh T, Gulrez G. Impact of immediate feedback on the learning of medical students in pharmacology. J Adv Med Educ Prof. 2019;7(1):1-6. DOI: 10.30476/JAMP.2019.41036

19. Chauhan GR, Chauhan BR, Vaza JV, Chauhan PR. Relations of the Number of Functioning Distractors With the Item Difficulty Index and the Item Discrimination Power in the Multiple Choice Questions. Cureus. 2023;15(7):e42492. DOI: 10.7759/cureus.42492

20. Rao C, Kishan Prasad H, Sajitha K, Permi H, Shetty J. Item analysis of multiple choice questions: Assessing an assessment tool in medical students. Int J Educ Psychol Res. 2016;2(4):201-204. DOI: 10.18203/2320-6012.ijrms20161256

21. Heist BS, Gonzalo JD, Durning S, Torre D, Elnicki DM. Exploring Clinical Reasoning Strategies and Test-Taking Behaviors During Clinical Vignette Style Multiple-Choice Examinations: A Mixed Methods Study. J Grad Med Educ. 2014;6(4):709-714. DOI: 10.4300/JGME-D-14-00176.1

22. Torre D, Daniel M, Ratcliffe T, Durning SJ, Holmboe E, Schuwirth L. Programmatic Assessment of Clinical Reasoning: New Opportunities to Meet an Ongoing Challenge. Teach Learn Med. 2025;37(3):403-411. DOI: 10.1080/10401334.2024.2333921

23. Croskerry P. Critical thinking and reasoning in emergency medicine. In: Croskerry P, Cosby KS, Schenkel SM, Wears RL, editors. Patient safety in emergency medicine. Philadelphia, PA: Lippincott Williams & Wilkins; 2008. p.213-218.

24. van Wijk EV, Janse RJ, Ruijter BN, Rohling JH, van der Kraan J, Crobach S, de Jonge M, de Beaufort AJ, Dekker FW, Lagers AM. Use of very short answer questions compared to multiple choice questions in undergraduate medical students: an external validation study. PLoS One. 2023;18(7):e0288558. DOI: 10.1371/journal.pone.0288558

25. Prakash S, Bihari S, Need P, Sprick C, Schuwirth L. Immersive high fidelity simulation of critically ill patients to study cognitive errors: a pilot study. BMC Med Educ. 2017;17(1):36. DOI: 10.1186/s12909-017-0871-x

26. De Beer M, Mårtensson L. Feedback on students' clinical reasoning skills during fieldwork education. Aust Occup Ther J. 2015;62(4):255-264. DOI: 10.1111/1440-1630.12208

Korrespondenzadresse:
Prof. Dr. med. Fabian Dupont, MHPE
Universität des Saarlandes, Zentrum Allgemeinmedizin, Geb. 80.2, 66424 Homburg (Saar), Deutschland
Fabian.dupont@uks.eu

Bitte zitieren als
Klutmann J, Dietzsch C, Schlasius-Ratter U, Oksche A, Volz-Willems S, Jordan S, Jäger J, Dupont F. Visualizing cognitive processes in medical education: Forward and backward reasoning in a digital family medicine assessment. GMS J Med Educ. 2026;43(5):Doc65. DOI: 10.3205/zma001859, URN: urn:nbn:de:0183-zma0018594

Artikel online frei zugänglich unter
https://doi.org/10.3205/zma001859

Eingereicht: *15.07.2025*
Überarbeitet: *21.11.2025*
Angenommen: *08.01.2026*
Veröffentlicht: *15.06.2026*