Leah T. Braun*, Katharina F. Borrmann, Christian Lottspeich, Daniel A. Heinrich, Jan Kiesewetter, Martin R. Fischer and Ralf Schmidmaier

Scaffolding clinical reasoning of medical students with virtual patients: effects on diagnostic accuracy, efficiency, and errors

<https://doi.org/10.1515/dx-2018-0090>

Received September 16, 2018; accepted January 21, 2019; previously published online February 18, 2019

Abstract

Background: Understanding clinical reasoning is a major challenge in medical education research. Little is known about the influence of scaffolding and feedback on the clinical reasoning of medical students. The aim of this study was to measure the effects of problem representation (cognitive representation of a clinical case) and structured scaffolding for reflection with or without feedback on the diagnostic efficiency and characterization of diagnostic errors of medical students.

Methods: One hundred and forty-eight advanced medical students were randomly assigned to one of five groups (2×2 design with a control group). They worked on 15 virtual clinical cases (five learning cases, five initial assessment cases, and five delayed assessment cases) in an electronic learning environment. After each case, they stated their presumed diagnosis and explained their diagnostic conclusion. Diagnostic accuracy, efficiency, and error distribution were analyzed.

Results: The diagnostic accuracy (number of correctly solved cases) and efficiency (solved cases/total time) did not differ significantly between any of the groups in the

two different assessment phases [mean = 2.2–3.3 (standard deviation [SD] = 0.79–1.31), $p = 0.08/0.27$ and mean = 0.07–0.12 (SD = 0.04–0.08), $p = 0.16/0.32$, respectively]. The most important causes for diagnostic errors were a lack of diagnostic skills (20%), a lack of knowledge (18%), and premature closure (17%).

Conclusions: Neither structured reflections nor representation scaffolding improved diagnostic accuracy or efficiency of medical students compared to a control group when working with virtual patients.

Keywords: clinical reasoning; diagnostic errors; scaffolding; undergraduate medical education.

Introduction

Inadequate clinical reasoning can lead to errors in daily clinical practice [1] and impact patient safety [1, 2]. In an earlier study, our group found that about 50% of diagnoses made by intermediate and advanced medical students working with virtual patients were incorrect [3]. Diagnostic errors made by medical students were divided into eight different categories [3]: errors caused by a lack of diagnostic skills, lack of knowledge, premature closure, misidentification, faulty context generation, faulty triggering, over- and underestimating, and no hypothesis. The first three were the most common errors.

To improve the clinical reasoning process of students and decrease the rate of misdiagnoses, different instructional approaches have been tested. Instructions that help students approach clinical cases are defined as scaffolding [4, 5]. Scaffolding during the diagnostic process seems to improve clinical reasoning, although the results are contradictory.

One type of scaffold is a representation prompt. Representations (or case summaries) can help students and experts to solve clinical cases [4, 6, 7], and practicing representations while diagnosing cases correlates with a correct solution [8]. In a previous study, we used representation prompts to foster clinical reasoning by medical students [4, 9]. Students were prompted to

*Corresponding author: Leah T. Braun, MD, Medizinische Klinik und Poliklinik IV, University Hospital, LMU Munich, Ziemssenstrasse 1, D-80336 Munich, Germany, Phone: +089-4400-57334, Fax: +089-4400-57339, E-mail: Leah.Braun@med.uni-muenchen.de

Katharina F. Borrmann, Jan Kiesewetter and Martin R. Fischer: Institute for Medical Education, University Hospital, LMU Munich, Munich, Germany, E-mail: Katharina.Borrmann@campus.lmu.de (K.F. Borrmann); jan.kiesewetter@med.uni-muenchen.de (J. Kiesewetter); martin.fischer@med.uni-muenchen.de (M.R. Fischer)

Christian Lottspeich, Daniel A. Heinrich and Ralf Schmidmaier: Medizinische Klinik und Poliklinik IV, University Hospital, LMU Munich, Munich, Germany, E-mail: Christian.Lottspeich@med.uni-muenchen.de (C. Lottspeich); daniel.heinrich@med.uni-muenchen.de (D.A. Heinrich); ralf.schmidmaier@med.uni-muenchen.de (R. Schmidmaier)

imagine a ward round setting and write down how they would present their patient to a colleague. Representation improved students' diagnostic efficiency but did not improve accuracy.

Structured reflection is another type of scaffold. One such approach is to ask participants to think about the three most-likely diagnoses in a case and to write down all the information that supports and refutes their diagnoses. They are also asked to think about information that is missing in the case but would support their differential diagnoses [10]. Structured reflection can foster the diagnostic accuracy of medical students (including at a 1-week delayed assessment) [10] but these results are not seen in all studies [11]. It is unknown how structured reflection might affect the time spent on a case and if the causes of diagnostic errors are influenced by structured reflection.

While there have been many studies on structured reflection and feedback (on diagnosis) and a few studies on representation, no data are available about the effects of representation scaffolds in combination with feedback. Also, the long-term effects of representation scaffolds have not been tested. Finally, there is no direct comparison of the effects of representation and reflection. Another important aspect in daily clinical practice is diagnostic efficiency. Due to limited time and resources in daily clinical practice, it is crucial to diagnose efficiently. Diagnostic efficiency can be defined as the number of (correctly) solved cases divided by the time needed for diagnosis [4]. Representation scaffolds can improve diagnostic efficiency in certain settings but it is unclear how feedback and structured reflection influence it.

Based on the literature and previous studies, we hypothesize the following:

1. Scaffolding with feedback is superior to scaffolding without feedback in improving diagnostic accuracy.
2. Representation scaffolding in a learning phase should lead to a more efficient diagnostic process in an immediate and delayed assessment than a reflection scaffold.
3. Reflection scaffolding is superior to representation scaffolding for improving diagnostic accuracy. The effects of structured reflection are still measurable with 1-week delay.
4. Both scaffolding methods (with or without feedback) are superior to a control group regarding diagnostic accuracy, efficiency, and rate of premature closure.

We chose a 2×2 design (representation or reflection scaffold with or without feedback) with a control group (no scaffold and no feedback) in a prospective, randomized, double-blinded setting to answer these questions.

Materials and methods

Participants

One hundred and fifty fourth and fifth year medical students from different medical schools in Germany who had successfully completed their courses in internal medicine participated in this study in June 2017. Participants received €30 for participating. Informed consent was obtained from all individuals included in this study. The study was approved by the Ethical Committee of the Medical Faculty of LMU Munich (No. 75-16).

Study design

Participants were randomly assigned to one of five groups (each $n=30$): a representation scaffold group (R0), a representation scaffold group with feedback (RF), a structured reflection group (S0), a structured reflection group with feedback (SF), and a control group (O0). All participants analyzed five learning cases except the control group which immediately started working on the assessment cases. After the learning phase, the four intervention groups completed assessment 1 (five cases) and after 1 week, assessment 2 (five cases). The dependent variables were: accuracy (binary variable of correct or incorrect diagnosis), time on task (tracked by the electronic learning platform CASUS) [12], and diagnostic efficiency (correct diagnoses divided by time spent on task).

The nature of diagnostic errors was assessed by coding students' written explanations and by tracking and analyzing the diagnostic steps as published previously [3]. One investigator (KB) coded all the explanations. A second rater (LB) coded 15% of the explanations. The interrater coefficient analyzed with Cohen's kappa was $k=0.88$. Both raters were blind to the group assignment. Both were familiar with the cases and with the coding scheme from the earlier study [3]. First, a prototypical error for all eight categories was defined. The raters read the explanations and any errors were assigned to one of eight categories. The categories are defined in the following:

Type	Definition [3, 13]
Knowledge base inadequate	Insufficient knowledge of relevant condition
Skills inadequate	Insufficient diagnostic skills for relevant condition
Faulty context generation	Lack of awareness of relevant aspects of the case
Overestimating/underestimating	Focus too closely on an aspect or failure to appreciate the relevance
Faulty triggering	Inappropriate conclusion
Misidentification	One diagnosis is mistaken for another
Premature closure	Failure to consider other possible diagnosis
No hypothesis	Failure to find any diagnosis at all

Virtual patients

Figure 1 shows the course of the study. First, students were randomly assigned to one of the five groups. After an introduction video (which explained how to proceed through the cases and how the diagnostic tests can be selected), the students completed a questionnaire regarding socio-demographic characteristics. They also completed a pre-test on knowledge about dyspnea (consisting of 15 single-choice questions). They then worked within the electronic case simulation platform CASUS [12] on 15 clinical cases. All cases were structured identically. They started with a history including the chief complaint, history of present illness, medical history, and medication and drug use. This was followed by a full and standardized physical examination. Subsequently, participants could freely select among 10 diagnostic tests from a patient record in each case. The case history was compatible with at least three diagnoses. The contents of each patient record supported one diagnosis but the records also purposely included incidental and distracting findings such as slightly elevated laboratory parameters. The patient record information was well-matched to the pre-existing conditions and comorbidities of the patients [e.g. poor lung function in a patient with chronic obstructive pulmonary disease (COPD) as pre-existing illness]. The time spent analyzing information was not restricted. Finally, participants were free to decide when to close the case by naming and explaining their final diagnosis. The cases were developed by the study authors LB and KB based on anonymized material from real cases and were reviewed by four physicians. Diagnostic accuracy was binary coded (correct or incorrect) according to an expert solution of the case that was determined in advance. In a pilot phase, 10 students tested the feasibility of the study procedure and assessed the difficulty of the cases.

Medical encounters

The cases in the learning phase pertained to dyspnea (see Figure 2). In assessment 1, two of the cases carried the identical diagnosis as in the

Learning phase	Assessment 1	Assessment 2
<ul style="list-style-type: none">• Asthma attack• Pneumothorax• COPD• Myocarditis• Tuberculosis• -----	<ul style="list-style-type: none">• Hyperventilation• Pneumothorax• COPD• Heart failure• -----• Cholelithiasis	<ul style="list-style-type: none">• Lung embolism• -----• COPD• Lung cancer• Tuberculosis• Diverticulitis

Figure 2: Medical content of the clinical cases.

learning phase. Another two of the cases shared possible differential diagnosis with the cases in the learning phase (myocarditis→heart failure, asthma→hyperventilation). The final case had abdominal pain as the primary symptom. In assessment 2, two cases had the same diagnosis as in the learning phase, two cases shared differential diagnoses, and one case had different content.

Scaffolding

In the representation scaffold (R) group, the CASUS system interrupted case processing after the medical history, physical exam, and initial test results with the following case representation prompt: “Please sum up the case in 1–2 sentences as you would present it to your attending expert.” [4] In the structured reflection (S) group, the case processing was also interrupted at the same stage. Students had to write down their leading diagnosis and their differential diagnoses and had to explain which information supported their diagnosis and which information was contrary to it [10].

Feedback

The feedback groups received a standardized feedback after giving their diagnosis in each case. The feedback was identically structured for all five cases: about 100 words long, explaining the important symptoms of the history and the key findings of the physical examination in addition to an explanation about the important diagnostic tests and important differential diagnoses (see Appendix). The no-feedback groups did not receive feedback or the correct solutions of the cases.

Statistics

In an *a priori* power analysis for chi-squared (χ^2) tests with $df=4$, we calculated that overall at least 100 participants should be tested in order to find a difference between the groups, assuming a small to medium effect (power: $1 - \beta=0.80$, Type I error: $\alpha=0.05$, Cohen’s $d=0.4$). Gaussian distribution was tested by the Kolmogorov-Smirnov test. Due to a lack of normal distribution, differences between groups were tested by the Kruskal-Wallis test. p-Values were adjusted with the Bonferroni method, and p-values of ≤ 0.006 were considered to indicate statistical significance.

Figure 1: Study design.

Results

Characterization of the participants and learning phase

One hundred and forty-eight of 150 participants (108 women) processed all cases and were included in the data analysis. Table 1 shows the characteristics of the study population. There were no statistically significant differences between the five groups' demographic variables. Prior knowledge with respect to dyspnea did not differ between the groups [$\chi^2(4)=4.81$, $p=0.31$].

The learning time (see Table 2) was not different between the groups [$\chi^2(3)=8.62$, $p=0.04$]: the representation groups needed mean=66.1 [standard deviation (SD)=20.9] min (without feedback) and mean=65.9 (SD=27.1) min (with feedback) to solve the five cases, whereas the reflection group needed mean=78.5 (SD=24.2) min (without feedback) and mean=62.3 (SD=24.9) min (with feedback). The participants solved between 2.7 and

3.1 of the five learning cases correctly; there was no significant difference between the groups [$\chi^2(3)=1.59$, $p=0.66$]. The diagnostic efficiency did not differ between the groups in the learning phase [$\chi^2(3)=2.97$, $p=0.40$]. The groups that received feedback spent between 30 and 60 s in each case reading the feedback.

Assessment: accuracy, efficiency, and diagnostic errors

Structured reflection is not superior to representation scaffold and is not superior to control (no scaffolding) with respect to accuracy (Table 3). All groups solved the same number of cases correctly [between 2.2 and 3.3 cases in the first assessment, $\chi^2(4)=8.40$, $p=0.08$]. The assessment cases that had the same diagnoses as the cases in the learning phase were not solved more often by the scaffolding groups. The accuracy in all 15 cases is shown in Table 4.

Table 1: Study participants' characteristics.

Group	S0	SF	R0	RF	00	p-Value
	M (SD) n=28	M (SD) n=30	M (SD) n=30	M (SD) n=30	M (SD) n=30	
Age, years	24.5 (2.5)	24.5 (2.5)	25.4 (3.7)	25.3 (3.2)	26.9 (4.0)	$\chi^2(4)=9.23$, $p=0.06$
Grade of school leaving examination ("Abitur") ^a	1.6 (0.7)	1.5 (0.5)	1.5 (0.4)	1.7 (0.6)	1.8 (0.7)	$\chi^2(4)=7.69$, $p=0.10$
National Boards Part 1 ^b after 2 years	2.1 (1.0)	2.1 (0.8)	2.3 (1.0)	2.6 (1.0)	2.5 (0.9)	$\chi^2(4)=5.12$, $p=0.28$
– Oral						
– Written	2.5 (0.9)	2.3 (0.9)	2.6 (0.9)	2.6 (1.0)	2.7 (0.9)	$\chi^2(4)=3.25$, $p=0.52$
Grade of internal medicine curriculum ^c	1.8 (1.1)	1.8 (1.2)	2.3 (1.2)	2.5 (1.4)	2.4 (1.5)	$\chi^2(4)=5.57$, $p=0.23$
Clerkships, months	3.3 (1.1)	3.4 (0.9)	3.2 (0.9)	3.2 (0.9)	3.3 (1.1)	$\chi^2(4)=1.42$, $p=0.84$
Score of conceptual knowledge pre-test (max. 15)	12.3 (2.8)	11.6 (2.6)	12.4 (2.8)	11.5 (3.3)	12.9 (2.7)	$\chi^2(4)=4.81$, $p=0.31$

R0, representation scaffold without feedback; RF, representation scaffold group with feedback; S0, structured reflection without feedback; SF, structured reflection group with feedback; 00, control group, p between groups. ^aSimilar to A-levels (1.0 best grade; 4.0 worst grade).

^bSimilar to USMLE-Step 1. ^cGrade in the internal medicine examination, which takes place in the third to fifth clinical year (depending on the university's curriculum).

Table 2: Learning phase.

	S0	SF	R0	RF	p-Value
	M (SD)	M (SD)	M (SD)	M (SD)	
Accuracy	3.1 (1.05)	2.7 (1.15)	2.9 (1.41)	3.0 (1.43)	$\chi^2(3)=1.59$, $p=0.66$
Efficiency	0.05 (0.02)	0.05 (0.03)	0.04 (0.02)	0.06 (0.04)	$\chi^2(3)=2.97$, $p=0.39$
Time on cases	66.1 (20.92)	65.9 (27.09)	78.5 (24.23)	62.3 (24.90)	$\chi^2(3)=8.62$, $p=0.04$
Number of diagnostic tests	44.5 (9.07)	45.5 (11.63)	45.7 (11.29)	55.5 (17.42)	$\chi^2(3)=6.19$, $p=0.10$

R0, representation scaffold without feedback; RF, representation scaffold group with feedback; S0, structured reflection without feedback; SF, structured reflection group with feedback; 00, control group. Efficiency=cases/total time. p-Value between groups.

Table 3: Assessment 1 (immediate).

	S0		SF		R0		RF		O0		p-Value
	M	(SD)	M	(SD)	M	(SD)	M	(SD)	M	(SD)	
Accuracy	2.5 (0.79)		2.6 (1.10)		2.4 (1.07)		2.2 (0.95)		3.0 (1.30)		$\chi^2(4) = 8.40, p = 0.08$
Efficiency	0.09 (0.06)		0.12 (0.08)		0.10 (0.07)		0.10 (0.06)		0.07 (0.04)		$\chi^2(4) = 6.54, p = 0.16$
Time	30.1 (12.39)		26.7 (17.58)		27.0 (11.35)		24.8 (9.82)		45.2 (14.45)		$\chi^2(4) = 40.24, p = 0.001$
Errors	30% skills		27% misidentification		27% skills		28% premature closure		19% underestimating		—
	20% misidentification		23% premature closure		22% underestimating		23% skills		17% premature closure		
	17% underestimating		17% skills		13% faulty context		20% misidentification		17% misidentification		
	11% premature closure		17% underestimating		13% premature closure		13% underestimating		17% skills		
	7% faulty context		7% faulty triggering		13% misidentification		8% faulty context		17% knowledge		
	6% faulty triggering		5% faulty triggering		10% no hypothesis		4% no hypothesis		4% faulty triggering		
	6% no hypothesis		3% no hypothesis		3% faulty triggering		3% knowledge		4% faulty context		
	4% a knowledge		2% knowledge		0% knowledge		1% faulty triggering		4% no hypothesis		

R0, representation scaffold without feedback; RF, representation scaffold group with feedback; S0, structured reflection group with feedback; SF, structured reflection group with feedback; O0, control group. Efficiency = cases/total time. p-Value between groups. Skills = lack of skills, underestimating = over- and underestimating, faulty context = faulty context generation, knowledge = a lack of knowledge.

Diagnostic efficiency did not differ between the groups [$\chi^2(4) = 6.54, p = 0.16$]. None of the groups worked more efficiently than the others. However, during assessment 1, time on task was significantly shorter in the interventional groups compared to the control group [$\chi^2(4) = 40.24, p < 0.001$] with a strong effect size of Cohen's $d = 1.16$. Whereas the control group needed 45 min (SD = 14), the other groups just needed between 24 and 30 min (SD = 10–18) (see Table 5).

The diagnostic error distribution was different between groups (Table 3). Participants misdiagnosed 814 out of 2080 cases. In the first assessment, the error distribution differed between the groups. A lack of skills was the most common error in the two scaffolding groups without feedback. In the groups with feedback, misidentification and premature closure were the most frequent causes for diagnostic errors. In the control group, multiple error types were equally frequent. The most common diagnostic errors with regard to the medical content are shown in Table 6.

In assessment 2 (after a 1-week interval from assessment 1), the accuracy was similar in all groups [$\chi^2(4) = 5.21, p = 0.27$] as was diagnostic efficiency [$\chi^2(4) = 4.70, p = 0.32$]. Regarding time on task, the differences seen with the first assessment were no longer apparent [$\chi^2(4) = 2.52, p = 0.64$]: All groups needed between 30 and 36 min (SD = 7–20) to solve the cases. In assessment 2, the most important errors in all five groups were a lack of knowledge and premature closure (Table 7).

Discussion

Result 1: Structured reflection is not superior to representation scaffold and not superior to control with respect to accuracy

While the degree of difficulty was sufficient to detect differences between the groups, diagnostic accuracy was similar across conditions. Neither representation nor structured reflections with or without feedback improved diagnostic accuracy. Why? There are many variables that can influence diagnostic accuracy and only a few of them are modified by scaffolding. To solve clinical cases, an adequate knowledge base [14], good diagnostic skills (e.g. interpretation of an electrocardiogram), and the correct interpretation and combination of clinical findings are obligatory [15]. The feedback did not improve diagnostic accuracy. It is possible that the feedback was too short, too standardized, or that the students did not pay enough attention to it. Feedback has not been consistently effective in improving diagnostic accuracy [16].

Table 4: Accuracy in all 15 cases.

Group	S0	SF	R0	RF	00	p-Value
	M (SD)	M (SD)	M (SD)	M (SD)	M (SD)	
Case 1	0.36 (0.49)	0.47 (0.51)	0.40 (0.50)	0.47 (0.51)	–	$\chi^2(3)=1.02, p=0.80$
Case 2	0.82 (0.39)	0.70 (0.47)	0.77 (0.43)	0.83 (0.38)	–	$\chi^2(3)=1.91, p=0.52$
Case 3	0.86 (0.36)	0.70 (0.47)	0.77 (0.43)	0.77 (0.43)	–	$\chi^2(3)=2.02, p=0.57$
Case 4	0.54 (0.51)	0.47 (0.51)	0.50 (0.51)	0.53 (0.51)	–	$\chi^2(3)=0.37, p=0.95$
Case 5	0.50 (0.51)	0.37 (0.49)	0.43 (0.50)	0.43 (0.50)	–	$\chi^2(3)=1.04, p=0.79$
Case 6	0.39 (0.50)	0.37 (0.49)	0.40 (0.50)	0.17 (0.38)	0.40 (0.50)	$\chi^2(4)=5.34, p=0.25$
Case 7	0.75 (0.44)	0.73 (0.45)	0.77 (0.43)	0.77 (0.43)	0.63 (0.49)	$\chi^2(4)=1.88, p=0.76$
Case 8	0.54 (0.51)	0.67 (0.48)	0.53 (0.51)	0.47 (0.51)	0.73 (0.45)	$\chi^2(4)=5.85, p=0.21$
Case 9	0.32 (0.48)	0.50 (0.51)	0.30 (0.47)	0.47 (0.51)	0.63 (0.49)	$\chi^2(4)=8.95, p=0.06$
Case 10	0.46 (0.51)	0.33 (0.48)	0.37 (0.49)	0.30 (0.47)	0.57 (0.50)	$\chi^2(4)=5.82, p=0.21$
Case 11	0.44 (0.51)	0.60 (0.50)	0.33 (0.48)	0.56 (0.51)	0.27 (0.45)	$\chi^2(4)=9.57, p=0.05$
Case 12	0.78 (0.42)	0.83 (0.38)	0.77 (0.43)	0.70 (0.47)	0.83 (0.38)	$\chi^2(4)=1.94, p=0.75$
Case 13	0.26 (0.45)	0.23 (0.43)	0.17 (0.38)	0.19 (0.40)	0.20 (0.41)	$\chi^2(4)=0.95, p=0.92$
Case 14	0.56 (0.51)	0.70 (0.47)	0.47 (0.51)	0.67 (0.48)	0.73 (0.45)	$\chi^2(4)=6.14, p=0.19$
Case 15	0.78 (0.42)	0.90 (0.31)	0.83 (0.38)	0.81 (0.40)	0.70 (0.47)	$\chi^2(4)=4.11, p=0.39$

Table 5: Time on task (in minutes) in all 15 cases.

Group	S0	SF	R0	RF	00	p-Value
	M (SD)	M (SD)	M (SD)	M (SD)	M (SD)	
Case 1	17.11 (7.01)	17.80 (8.13)	22.23 (10.03)	16.97 (12.11)	–	$\chi^2(3)=8.22, p=0.04$
Case 2	11.61 (4.44)	12.13 (6.45)	15.43 (5.46)	10.77 (4.39)	–	$\chi^2(3)=14.26, p=0.003$
Case 3	12.25 (4.59)	12.17 (6.18)	14.83 (5.34)	11.50 (4.56)	–	$\chi^2(3)=8.40, p=0.04$
Case 4	12.11 (4.45)	11.47 (5.08)	13.47 (3.61)	10.93 (4.52)	–	$\chi^2(3)=8.58, p=0.04$
Case 5	13.04 (5.34)	12.33 (5.97)	12.57 (5.38)	12.17 (4.21)	–	$\chi^2(3)=0.74, p=0.86$
Case 6	7.04 (4.44)	5.60 (3.85)	6.70 (6.43)	6.43 (3.88)	12.67 (5.42)	$\chi^2(4)=39.54, p=0.001$
Case 7	5.57 (2.17)	5.40 (3.92)	5.10 (3.39)	5.23 (2.11)	10.43 (4.29)	$\chi^2(4)=38.69, p=0.001$
Case 8	5.96 (3.72)	4.83 (3.06)	5.37 (2.85)	4.10 (1.88)	7.80 (3.67)	$\chi^2(4)=22.98, p=0.001$
Case 9	6.25 (3.10)	5.27 (4.56)	4.77 (2.80)	4.40 (2.33)	7.30 (4.82)	$\chi^2(4)=13.06, p=0.01$
Case 10	5.29 (2.48)	5.60 (4.24)	5.07 (2.35)	4.63 (1.79)	7.00 (3.26)	$\chi^2(4)=12.13, p=0.02$
Case 11	7.48 (2.95)	8.07 (4.75)	9.07 (4.76)	7.00 (2.50)	8.67 (3.61)	$\chi^2(4)=3.63, p=0.46$
Case 12	6.63 (2.79)	6.97 (3.78)	7.67 (3.75)	7.00 (3.09)	6.60 (2.11)	$\chi^2(4)=1.50, p=0.83$
Case 13	7.78 (5.07)	8.83 (6.23)	7.77 (5.57)	6.44 (2.91)	7.40 (2.53)	$\chi^2(4)=4.25, p=0.37$
Case 14	5.37 (2.44)	7.00 (5.49)	6.43 (3.40)	4.74 (1.68)	5.50 (2.33)	$\chi^2(4)=3.64, p=0.46$
Case 15	5.07 (2.34)	5.27 (2.64)	5.37 (3.69)	4.96 (2.59)	4.80 (2.14)	$\chi^2(4)=0.79, p=0.94$

Representation scaffolds and structured reflections did not improve diagnostic accuracy in the case scenarios we used. This corresponds to findings from other studies, which showed that groups which received instructional interventions are not more accurate than those that wrote down their first diagnostic idea [11]. In a previous study, representation did not improve accuracy but did improve the diagnostic efficiency of medical students [4]. The different results compared to another study [10] which studied the effects of reflection might be explained by a different degree of complexity of the cases. Our cases were quite complex, as they included 10 different diagnostic tests and distractors. In addition, our participants received the structured reflection prompt twice during case processing, while in the other study, the students used structured reflection only once at the end of the case.

Result 2: Representation scaffold is not superior to structured reflection and control with respect to efficiency

The time spent on the assessment cases was shorter in the scaffolding groups, but the shorter time on task did not translate into less accuracy. It has been shown that time pressure leads to less diagnostic accuracy [17], but simply solving cases faster without external pressure to do so does not cause more diagnostic errors.

The habituation to the learning environment in the learning phase may explain the shorter time on case in the scaffolding groups. The control group was as fast as the other groups in the second assessment. It is possible that it is not the interventions themselves but rather the learning with cases is helpful to decrease the time spent on a case.

Table 6: Most common diagnostic errors with regard to the medical content.

Case	Group 1 (S0)	Group 2 (SF)	Group 3 (R0)	Group 4 (RF)	Group 5 (00)
Tuberculosis 1	Lack of Knowledge	Lack of knowledge	Lack of knowledge	Lack of knowledge	–
Pneumothorax 1	Lack of skills	Lack of skills	Lack of skills	Lack of skills	–
COPD 1	Context generation	Overestimating	Lack of skills	Lack of knowledge	–
Myocarditis	Lack of skills	Misidentification	Misidentification	Misidentification	–
Asthma attack	Premature closure	Premature closure	Premature closure	Premature closure	–
Hyperventilation	Context generation	Premature closure	Overestimating	Premature closure	Overestimating
Heart failure	Diverse	Misidentification	Misidentification	Misidentification	Lack of knowledge
COPD 2	Lack of skills	Lack of skills	Lack of skills	Lack of skills	Lack of skills
Pneumothorax 2	Lack of skills	Lack of skills/overestimating	Lack of skill/overestimating	Lack of skills, overestimating	Lack of skills, overestimating
Cholelithiasis	Misidentification	Misidentification	Misidentification	Misidentification	Misidentification
Tuberculosis 2	Lack of skills, premature closure	Premature closure	Lack of knowledge	Faulty context	Lack of knowledge, faulty context generation
Lung embolism	Lack of knowledge	Lack of knowledge	Lack of knowledge, overestimating	Lack of knowledge	Diverse
Lung cancer	Lack of skills	Lack of skills	Premature closure	Lack of skills, premature closure	Lack of skills
COPD 3	Premature closure	Premature closure	Diverse	Diverse	Premature closure
Diverticulitis	Lack of knowledge	Lack of knowledge	Lack of knowledge, misidentification	Lack of knowledge	Lack of knowledge

Table 7: Assessment 2 (1 week delayed).

Group	S0	SF	R0	RF	00	p-Value
	M (SD)	M (SD)	M (SD)	M (SD)	M (SD)	
Accuracy	2.8 (1.18)	3.3 (0.98)	2.6 (1.31)	2.9 (1.41)	2.7 (1.17)	$\chi^2 (4) = 5.21, p = 0.27$
Efficiency	0.09 (0.05)	0.11 (0.07)	0.08 (0.06)	0.11 (0.07)	0.09 (0.04)	$\chi^2 (4) = 4.70, p = 0.32$
Time	32.3 (10.68)	36.1 (19.80)	36.3 (16.56)	30.2 (7.98)	33.0 (8.67)	$\chi^2 (4) = 2.52, p = 0.64$
Errors	27% skills	29% knowledge	29% knowledge	24% knowledge	23% premature closure	–
	25% knowledge	25% premature closure	17% premature closure	19% premature closure	23% knowledge	
	20% premature closure	21% skills	15% faulty context	19% faulty context	20% skills	
	12% faulty context	11% faulty context	12% skills	15% skills	13% faulty context	
	8% misidentification	7% underestimating	12% underestimating	9% underestimating	10% underestimating	
	4% underestimating	5% misidentification	8% misidentification	7% misidentification	7% no hypothesis	
	2% faulty triggering	2% no hypothesis	3% faulty triggering	6% faulty triggering	3% misidentification	
	2% no hypothesis	0% faulty triggering	3% no hypothesis	2% no hypothesis	2% faulty triggering	

R0, representation scaffold without feedback; RF, representation scaffold group with feedback; S0, structured reflection without feedback; SF, structured reflection group with feedback; 00, control group Efficiency = cases/total time. p-Value between groups. Skills = lack of skills, underestimating = over- and underestimating, faulty context = faulty context generation, knowledge = a lack of knowledge.

Result 3: Diagnostic error distribution is different between groups

The distribution of the diagnostic errors was modified by the scaffolding in the first assessment. Groups that received feedback made more errors due to premature closure and misidentification. It is possible that the feedback led to increased confidence so that the students felt quite sure about their diagnoses in the assessments. Blissett and Sibbald pointed out that “collecting limited information may be described as efficient when the diagnosis is correct and as PCB (premature closure bias) when the diagnosis is incorrect” [18]. We could not show an association between premature closure and time on task or efficiency in our study. The groups that received scaffolding and feedback misdiagnosed the most due to premature closure. In our study, students tend to collect almost all of the offered information, but information that is contradictory to their initial diagnosis cannot correct their diagnostic conclusions.

In the second assessment, the scaffolding did not influence the distribution of errors but specific cases were associated with certain kinds of errors. Diagnostic errors are at least partly content-dependent [3]. This finding might be useful for further studies. If certain cases are associated with certain errors (for example, premature closure), those cases could be used for teaching purposes. Similar to the time spent on the cases, the effects on diagnostic error distribution were not reproduced in the second assessment.

Limitations

As we used a laboratory setting with only one short intervention, we cannot comment on the transfer into educational field conditions. The feedback may not have been extensive enough to improve diagnostic performance. These scaffolding methods might be helpful in other expertise levels [19].

Conclusions and outlook

This is the first study that compared representation and reflection scaffolds and the influence of feedback on clinical reasoning in comparison to a control group.

All interventions failed to improve accuracy and efficiency in immediate and delayed assessment. This

is an important finding: scaffolding methods, which have been successful in certain cases and settings, are not automatically effective in cases of a different complexity. The efficacy of scaffolding is likely influenced by the learner and the case. There is a lack of knowledge on how to adapt scaffolding to these circumstances. Prospective studies are needed to observe the influence of long-lasting courses with scaffolding. It remains unclear if the causes for diagnostic errors can be influenced by scaffolding.

In further studies, longer interventions and cases of differing complexity should be utilized.

In summary, in this randomized, controlled laboratory study of clinical reasoning in fourth and fifth year medical students, representation and reflection scaffolds did not improve diagnostic accuracy or reduce premature closure.

Acknowledgments: We are thankful to Laura Handgriff (Medizinische Klinik und Poliklinik IV, Klinikum der Universität München, LMU) and Peter Musaeus (Centre for Health Sciences Education, Aarhus, Denmark) for critically proofreading the manuscript.

Author contributions: LB contributed to the conceptual design of the study, the analysis and interpretation of the data, and the drafting of the manuscript. KB contributed to the study design and the collection of data and substantially contributed to the drafting of the manuscript. CL and DH substantially contributed to the study material and the drafting of the manuscript. JK contributed to the design of the study and the revision of the manuscript. MRF contributed to the conceptual design of the study, the interpretation of data, and the revision of the paper. RS contributed to the conceptual design of the study, the collection, analysis, and interpretation of data, and the drafting and revision of the paper. All the authors have accepted responsibility for the entire content of this submitted manuscript and approved submission.

Research funding: Funds for this project were provided by the Friedrich-Baur-Institut, University Hospital, LMU Munich.

Employment or leadership: None declared.

Honorarium: None declared.

Competing interests: The funding organization(s) played no role in the study design; in the collection, analysis, and interpretation of data; in the writing of the report; or in the decision to submit the report for publication.

Appendix

Feedback example

Anamnesis: Here it is important to pay attention to existing B-symptoms. There are also various risk factors for tuberculosis: immunosuppression, old age, travel to certain countries or the country of origin (Eastern Europe). Immunosuppressed patients are prone to reactivation of tuberculosis. The chronic course excludes acute pneumonia.

PE: Particular attention should be paid to the body temperature (sub-febrile) and the basally damped knocking sound, which suggests an infiltrate.

Patient record: In this case, an x-ray should be taken (DD: lung carcinoma). One can see here a miliary tuberculosis, which can occur especially in immuno-compromised patients. The lab indicates an infection. With the help of bacteriological examination the pathogen can be detected. The positive Ziehl-Neelsen stain is typical of mycobacteria.

Case example

The 61-year-old Mr. Hartmann complains of morning cough with whitish ejecta for a few months. In addition, he suffered from low breathlessness during physical exertion. Fever and night sweats are denied. Pre-existing conditions include hypertension treated with ramipril and a thiazide diuretic and simvastatin-treated hypercholesterolemia.

Diagnostic tests

Chest X-ray

He regularly drinks alcohol (2 beers/week) and has been smoking for over 40 years (about 60 packs per year).

In the physical examination, you make the following findings:

Awake, fully oriented, cognitively inconspicuous 61-year-old patient in good general condition and adipose nutritional status (100 kg, 1.86 m, BMI 30 kg/m²), affectionate and friendly, adequate verbal reaction.

Vital signs: blood pressure 150/95 mmHg, heart rate 82/min, respiratory rate 17/min, body temperature (ear) 36.6°C

Cardiovascular system: Heart sounds pure, rhythmic, no heart noises, no auscultated flow noises on the carotid. No jugular venous stasis, subtle peripheral edema. Peripheral pulse (A. radialis, A. dorsalis pedis, A. tibialis posterior) palpable.

Respiratory system: No chest deformities. Slight lip cyanosis. Hypersonic knocking sound on both sides, depending on respiration, 2 cm displaceable lung borders. Quiet breathing, bronchial breathing, no wet or dry secondary breath sounds. Thoracic spine not bumpy.

Abdomen: inspection unobtrusive. Bowel sounds vividly over all four quadrants, no abdominal flow noises. Soft abdominal wall, not painful palpation, no palpable resistance. Percussion normal knocking sound. Liver at inspiration with soft margin of the liver and without pain 3 cm caudal to the costal arch palpable, percussion 11 cm in the MCL. Spleen not palpable. No hernias. No kidney pain. Lumbar spine not bumpy.

Lymph nodes: no pathologically enlarged cervical, axillary, or inguinal lymph nodes palpable.

ECG

Lab tests

Erythrocyte	4.8 L/pL	4.2–5.1/pL
Hemoglobin	14.0 g/dL	12.0–16.0 g/dL
Hematocrit	0.43	0.36–0.46
Leukocytes	7.0 G/L	4.0–11.0 G/L
Thrombocytes	220 G/L	150–440 G/L
Creatinine	0.8 mg/dL	0.5–1.0 mg/dL
Sodium	139 mmol/L	135–145 mmol/L
Potassium	4.5 mmol/L	3.5–5.0 mmol/L
CRP	0.3 mg/dL	<0.5 mg/dL
SGOT	52 U/L	<33 U/L
SGPT	74 U/L	<35 U/L
γ-GT	45 U/L	<38 U/L
Quick	110%	70–130%
INR	0.9	0.8–1.2
aPTT	28 s	23–35 s
Glucose	152 mg/dL	70–115 mg/dL
Urea	6.0 mmol/L	2.0–8.0 mmol/L
GFR	>60	>60
TSH	1.5 mU/L	0.4–2.5 mU/L
Chloride	98 mmol/L	96–110 mmol/L
Magnesium	1.05 mmol/L	0.70–1.20 mmol/L
Calcium	2.41 mmol/L	2.15–2.60 mmol/L
BSG	12 mm/h	Between 15 mm and 20 mm/h
Bilirubin	0.9 mg/dL	<1.0 mg/dL
Cholesterol/TAG	280 mg/dL	150–220 mg/dL
Ferritin	110 µg/L	23–217 µg/L
CK-MB	3 µg/L	<5 µg/L
Troponin T	0.08 ng/mL	<0.1 ng/mL
Uric acid	6.4 mg/dL	2.5–7.0 mg/dL
D-Dimer	220 ng/mL	<500 ng/mL
PCT	0.02 µg/L	<0.05 µg/L
HbA _{1c}	8.2%	3.0–6.0%
Lactate	2.52 mmol/L	0.63–2.44 mmol/L
Differential:		
Neutrophils	66%	40–75%
Eosinophils	4%	1–6%
Basophiles	0%	0–1%
Monocytes	2%	2–8%
Lymphocytes	28%	20–45%

Neurological consultation

Awake, fully oriented, cognitively inconspicuous 61-year-old patient with adequate verbal response.

No focal neurological deficit recognizable.

Pupils: PERLA

Motor skills: no latent or overt paresis with normal muscle tone

Reflexes: equally medically releasable (PSR, ASR, BSR) without widened reflex zones, Babinski negative

Sensitivity: intact in all qualities

Cranial nerves: cranial nerves intact, oculomotor unremarkable, no nystagmus

Gear and coordination: age-appropriate unremarkable

Cognition: MMST 30/30

Blood gas analysis

pH	7.42	7.36–7.44
pO ₂	58 mmHg	75–98 mmHg
saO ₂	92%	95–97%
pCO ₂	50 mmHg	35–45 mmHg
Standard bicarbonate	24	20–28 mmol/L
BE	–1	–2 to +2 mmol/L
Sample type	Arterial	
FiO ₂	Room air	
Respiratory rate	16	

Lung function test

	Set	Actual	% (Actual/set)
VC MAX (L)	4.27	3.5	81.9
FEV 1 (L)	3.19	1.35	42.3
FEV 1% VC MAX (%)	75.51	38.5	50.1
MEF 50 (L/s)	4.31	2.25	42.2
MEF 25 (L/s)	1.56	0.78	50.0
TLC (L)	6.98	7.15	102.0
RV (L)	2.51	3.65	145.4
TLCOc/V A (mmol/min/kPa/L)	1.32	1.02	77.27

VC, vital capacity; FEV, forced expiratory volume; MEF, maximum expiratory flow; TLC, total lung capacity; RV, residual volume; TLCOc/V A, CO, transfer coefficient.

Echocardiography

Aortic root normal wide. AK: age-related, opening movement normal, no insufficiency, no stenosis. RA: moderately enlarged (visually assessed). LA: moderately enlarged. RV: normal size, walls moderately thickened,

normal RV function. LV: normal size, walls moderately thickened. Systolic global function low-normal (planimetric EF ~55%). No reg. Contraction disorders. Diastolic function normal. MK: age-related, movement normal, low insufficiency, no stenosis. TK: age-related, movement normal, low insufficiency, dpmax RV/RA = 25 mmHg. VCI not dilated, breath modulated.

Summary: LVEF low-normal, visually assessed moderately enlarged RA, no relevant vitals, normal diastolic function.

Sonography

Liver: image quality good. Size, normal, surface: smooth, echo pattern: condensed, no sound attenuation, vessel structures: normal, low-grade fat deposits (ICD-K76.0).

Gallbladder: well assessable, orthotopic, inconspicuous wall conditions, intraluminally no pathological echo structures, normal organ size.

Pancreas: body can be assessed with limited sensitivity, head and tail cannot be assessed during meteorism. Shape/contour: smooth, echo pattern: hypoechoic, pancreatic duct unrepresentable.

Spleen: Well assessable, normal size, echonormal homogenous internal reflex pattern, no focal changes, hilum free. Organ size: 105 mm × 59 mm × 42 mm, volume 136 mL.

Right kidney: easy to assess, orthotopic position, normal organ size, normal shape, age-related parenchyma, smooth organ contour, no urinary retention, no concrements. Size: 93 mm × 55 mm × 40 mm, volume 107 mL.

Left kidney: Well assessable, orthotopic position, normal organ size, normal shape, age-related parenchyma, smooth organ contour, no urinary retention, no concrements. Size: 94 mm × 50 mm × 37 mm, volume 91 mL.

Bladder: Well assessable, orthotopic, inconspicuous wall conditions, intraluminal no pathological echo structures, normal organ size.

Peritoneum: no ascites.

Prostate: normal findings.

Inconspicuous findings of the following abdominal organs: abdominal. Lymph nodes. Not assessable were: biliary tract.

Summary Assessment

Liver: low-grade fat deposits (ICD-K76.0)

Pancreas in strong air overlay only partially visible in the body. Other sonographic findings inconspicuous.

Urine analysis

Leukocytes	Negative
Nitrite	Negative
Protein	Negative
pH	6.5
Blood	Negative
Ketones	Negative
Glucose	Negative

Bacteriological tests

1. Material: sputum microscopy

Gram-positive cocci: negative	Erythrocytes: negative
Detritus: negative	Leukocytes: negative
Oral flora: +++	

Evaluation: negative microbiological findings

2. Material: blood culture

Evaluation: negative microbiological findings

Amounts are: +++ a lot, ++ moderate amount, + little, (+) very little

References

1. Graber ML, Carlson B. Diagnostic error: the hidden epidemic. *Physician Exec* 2011;37:12–8.
2. Graber ML. The incidence of diagnostic error in medicine. *BMJ Qual Saf* 2013;22(Suppl 2):ii21–7.
3. Braun LT, Zwaan L, Kiesewetter J, Fischer MR, Schmidmaier R. Diagnostic errors by medical students: results of a prospective qualitative study. *BMC Med Educ* 2017;17:191.
4. Braun LT, Zottmann JM, Adolf C, Lottspeich C, Then C, Wirth S, et al. Representation scaffolds improve diagnostic efficiency in medical students. *Med Educ* 2017;51:1118–26.
5. Benson BK. Coming to terms: scaffolding. *Engl J* 1997; 86:126–7.
6. Bordage G, Connell KJ, Chang RW, Gecht MR, Sinacore JM. Assessing the semantic content of clinical case presentations: studies of reliability and concurrent validity. *Acad Med* 1997;72(10 Suppl 1):S37–9.
7. Nendaz MR, Bordage G. Promoting diagnostic problem representation. *Med Educ* 2002;36:760–6.
8. Kiesewetter J, Ebersbach R, Görlitz A, Holzer M, Fischer MR, Schmidmaier R. Cognitive problem solving patterns of medical students correlate with success in diagnostic case solutions. *PLoS One*. 2013;8:e71486.
9. Braun LT, Lenzer B, Kiesewetter J, Fischer MR, Schmidmaier R. How case representations of medical students change during

- case processing – results of a qualitative study. *J Med Educ* 2018;35:Doc41.
10. Mamede S, Van Gog T, Sampaio AM, De Faria RMD, Maria JP, Schmidt HG. How can students' diagnostic competence benefit most from practice with clinical cases? The effects of structured reflection on future diagnosis of the same and novel diseases. *Acad Med* 2014;89:121–7.
 11. Ilgen JS, Bowen JL, McIntyre LA, Banh KV, Barnes D, Coates WC, et al. Comparing diagnostic performance and the utility of clinical vignette-based assessment under testing conditions designed to encourage either automatic or analytic thought. *Acad Med* 2013;88:1545–51.
 12. Fischer MR, Aulinger B, Baehring T. Computer-based-training (CBT): fallorientiertes lernen am PC mit dem CASUS/ProMediWeb-system. *Deut Med Wochenschrif* 1999;124:1401.
 13. Graber ML, Franklin N, Gordon R. Diagnostic error in internal medicine. *Arch Intern Med* 2005;165:1493–9.
 14. Schmidt HG, Rikers RM. How expertise develops in medicine: knowledge encapsulation and illness script formation. *Med Educ* 2007;41:1133–9.
 15. Elstein AS, Shulman LS, Sprafka SA. Medical problem solving an analysis of clinical reasoning. Cambridge: Harvard University Press; 1978.
 16. Lechermeier J, Fassnacht M. How do performance feedback characteristics influence recipients' reactions? A state-of-the-art review on feedback source, timing, and valence effects. *Manag Rev Q* 2018;68:145–93.
 17. ALQahtani DA, Rotgans JI, Mamede S, Mahzari MM, Al-Ghamdi GA, Schmidt HG. Factors underlying suboptimal diagnostic performance in physicians under time pressure. *Med Educ* 2018;52:1288–98.
 18. Blissett S, Sibbald M. Closing in on premature closure bias. *Med Educ* 2017;51:1095–6.
 19. Mamede S, Schmidt HG, Penaforte JC. Effects of reflective practice on the accuracy of medical diagnoses. *Med Educ* 2008;42:468–75.