MEDICAL TEACHER
An International Journal of Education in the Health Sciences
amee

Medical Teacher

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/imte20

Teaching medical students to apply deliberate reflection

Josepha Kuhn, Silvia Mamede, Pieter van den Berg, Laura Zwaan, Gijs Elshout, Patrick Bindels & Tamara van Gog

To cite this article: Josepha Kuhn, Silvia Mamede, Pieter van den Berg, Laura Zwaan, Gijs Elshout, Patrick Bindels & Tamara van Gog (2024) Teaching medical students to apply deliberate reflection, Medical Teacher, 46:1, 65-72, DOI: 10.1080/0142159X.2023.2229504

To link to this article: https://doi.org/10.1080/0142159X.2023.2229504

MEDICAL TEACHER
2024, VOL. 46, NO. 1, 65–72
https://doi.org/10.1080/0142159X.2023.2229504

MEDICAL TEACHER
Taylor & Francis
Taylor & Francis Group

Teaching medical students to apply deliberate reflection

Josepha Kuhn[a,b], Silvia Mamede[b,c], Pieter van den Berg[a], Laura Zwaan[b], Gijs Elshout[a], Patrick Bindels[a] and Tamara van Gog[d]

[a]Department of General Practice, Erasmus Medical Centre, Rotterdam, The Netherlands; [b]Institute of Medical Education Research Rotterdam, Erasmus Medical Centre, Rotterdam, The Netherlands; [c]Department of Psychology, Education and Child Studies, Erasmus University Rotterdam, Rotterdam, The Netherlands; [d]Department of Education, Utrecht University, Utrecht, The Netherlands

ABSTRACT

Purpose: *Deliberate reflection* on initial diagnosis has been found to repair diagnostic errors. We investigated the effectiveness of teaching students to use deliberate reflection on future cases and whether their usage would depend on their perception of case difficulty.
Method: One-hundred-nineteen medical students solved cases either with deliberate-reflection or without instructions to reflect. One week later, all participants solved six cases, each with two equally likely diagnoses, but some symptoms in the case were associated with only one of the diagnoses (*discriminating features*). Participants provided one diagnosis and subsequently wrote down everything they remembered from it. After the first three cases, they were told that the next three would be difficult cases. Reflection was measured by the proportion of discriminating features recalled (overall; related to their provided diagnosis; related to alternative diagnosis).
Results: The deliberate-reflection condition recalled more features for the *alternative* diagnosis than the control condition ($p = .013$) regardless of described difficulty. They also recalled more features related to their *provided* diagnosis on the first three cases ($p = .004$), but on the last three cases (described as difficult), there was no difference.
Conclusion: Learning deliberate reflection helped students engage in more reflective reasoning when solving future cases.

KEYWORDS
Deliberate reflection; diagnostic reasoning; diagnostic error; learning-by-teaching

Introduction

The deliberate-reflection procedure (Mamede et al. 2008) has been found to be an effective and consistently successful cognitive intervention to improve diagnostic accuracy (Lambe et al. 2016; Prakash et al. 2019), especially when physicians solve complex cases (Mamede et al. 2008) or are misled by cognitive bias (e.g. when physicians are distracted by a recently seen case that resembles the case at hand but has a different diagnosis, i.e. availability bias (Mamede et al. 2010), or by a patient's disruptive behaviour (Schmidt et al. 2017)). As physicians' first impression of a case influences how the presented information is interpreted, relevant features related to an alternative diagnosis may sometimes remain unnoticed (due to anchoring and confirmation bias (Wallsten 1981; Kostopoulou et al. 2012)). In these situations, it can help to go back to the case and analytically evaluate one's first impression. With deliberate reflection, physicians are asked to follow specific steps to systematically analyse multiple possible diagnoses for the case and how they relate to the findings from the case, before coming to a final conclusion. These steps aim to help physicians out of a tunnel vision induced by the first hypothesis, to sufficiently consider alternative diagnoses, and to correct initial mistakes.

Prior studies on the use of deliberate reflection have mainly focussed on the effect that it has on the diagnostic accuracy on a case (Mamede et al. 2008; Mamede et al. 2010, 2012; Schmidt et al. 2017; Costa Filho et al. 2019). A major open question is whether the procedure itself can be learned and then applied autonomously (i.e. without being prompted) when encountering future cases. If physicians would learn this in their medical training, it may help them to avoid some diagnostic errors in practice later on and may improve patient safety. While previous studies (Ibiapina et al. 2014; Mamede et al. 2019)

Practice points

- Learning deliberate reflection helped students engage in more reflective reasoning when solving future cases, regardless of described difficulty.
- Students who had not been taught deliberate reflection remembered more discriminating features (i.e. engaged in more reflective reasoning) when they expected cases to be difficult compared to cases that had not been described as difficult.
- Future studies should investigate whether teaching medical students the deliberate reflection procedure would also lead to improved diagnostic accuracy.

CONTACT Josepha Kuhn j.kuhn@erasmusmc.nl, j.k.g.kuhn@gmail.com Department of General Practice, Erasmus MC, P.O. Box 2040, 3000 CA Rotterdam, The Netherlands

Supplemental data for this article can be accessed online at https://doi.org/10.1080/0142159X.2023.2229504.

found that studying examples of deliberate reflection positively affected students' knowledge of the diagnoses shown in the examples, we cannot infer that they also learned and applied the deliberate reflection procedure itself. In a recent study with residents in general-practice training (Kuhn et al. 2021), we started to investigate whether learning deliberate reflection *via* a learning-by-teaching approach, would affect future diagnostic reasoning. In that study, participants who practised with deliberate reflection first studied examples that showed how the procedure was used (i.e. deliberate-reflection models). After that, participants explained what they had learned to a fictitious peer while being video recorded. Learning by teaching fosters an active engagement with the material, and adding the video camera induces more arousal, which have both been found to improve learning (Van Gog and Rummel 2010; Hoogerheide et al. 2019). The recorded videos of our previous study showed that participants in the learning-by-teaching condition had indeed learned the steps of the deliberate-reflection procedure. On a delayed test a couple of days later, all participants diagnosed new cases while thinking aloud. Against our expectations, participants in the deliberate-reflection condition did not show more elements of the learned procedure in their think-aloud protocols than did the control condition.

We consider three possible explanations for these results. The first one is that even though participants had learned the deliberate-reflection procedure, they did not apply it in the test phase because they did not feel the need to. We know that participants engage in more reflective reasoning when cases are described as being difficult (Mamede et al. 2008). A second explanation is that the think-aloud task, which was used to measure the residents' reasoning in the test-phase, evoked a more analytical approach for all residents (including the control condition). A post-hoc measurement of reflective reasoning may be better in order not to influence participants' reasoning during the diagnostic task. A third explanation is, that residents are already too experienced in diagnosing cases and therefore do not easily adopt a new way of diagnostic reasoning. Less experienced physicians in training, like medical students, may adopt deliberate reflection more easily.

Therefore, in the present study, we investigated whether medical students would learn the deliberate-reflection procedure by first studying an example and then explaining it to a fictitious peer (learning session) and whether this would influence their reasoning in novel cases one week later (test session) when compared to a control condition. In the test session, we used a recall task after diagnosing to measure the participants' reflective reasoning, which has been used for this purpose in previous studies (Mamede et al. 2007). We only focussed on recalled symptoms that helped to discriminate between possible diagnoses. Engaging in deliberate reflection requires analysing and weighing several diagnoses. Therefore, we expected that when students engaged in deliberate reflection, they would recall more of the relevant features related to not only their own but also the alternative diagnosis, as focussing on them may help to avoid a tunnel vision based on their first impression of the case. A reflective approach may also be reflected in more time spent diagnosing a case (Mamede et al. 2007) and participants may report more mental effort investment than with a non-analytical approach (Ibiapina et al. 2014; Mamede et al. 2019). Furthermore, we expected that the effect of practising with deliberate reflection would be more pronounced or only show when the cases in the test phase had been described as being difficult than when no description of difficulty was given, even if the difficulty of the cases does not actually change, because students may only feel the need to apply deliberate reflection when cases are expected to be difficult (Mamede et al. 2008).

Method

Design

The study consisted of a learning phase and a test phase (Figure 1). During the learning phase, twenty pre-existing student groups were randomly assigned to either the deliberate-reflection condition, where they studied examples and then explained the deliberate-reflection procedure to a fictitious peer, or to the control condition where they diagnosed cases without further instructions. About one week later, all students took the same test on six new, clinical cases of which three were described as being difficult cases even though the difficulty did not actually change. They were asked to first diagnose a case and then complete a free recall activity by writing down everything they remembered from the case. The ethics committee of the Erasmus Medical Centre viewed the research proposal and granted exemption from further review. The authors report there are no competing interests to declare.

Participants

In 2019, we invited 138 medical students from the Erasmus Medical Centre in Rotterdam who followed the general-practice track of the clerkships in the fifth or sixth year of their basic medical training of which 124 attended and completed both sessions. Three participants were removed for not filling in the informed consent form, and two more participants were excluded for doing the test session earlier than five or later than nine days after the learning sessions, which left us with a final data set of 119 participants (74 female, 45 male; age $M = 25.41$, $SD = 1.78$). One of the participants did not state their age. Supplementary Table 1 presents (additional) demographic information separately for each study condition. The study took place during the usual educational program. Every two weeks, one or two pre-existing student groups would participate in the study. A group consisted of five to twelve students. We alternated between assigning a group to the control condition or the deliberate reflection condition. To stimulate participation and compensate for the invested time, students who participated in this study could skip an assignment of their usual educational program.

At the Erasmus Medical Centre, students are trained in clinical reasoning in the Bachelor and Master medical programme. During the three years of the Bachelor, there are twelve lectures specifically concerning clinical reasoning and 12 small-group sessions (with twelve to fourteen students in each group). In the small-group sessions, students are presented a clinical case and trained in the process of clinical reasoning under the guidance of a clinician. In the

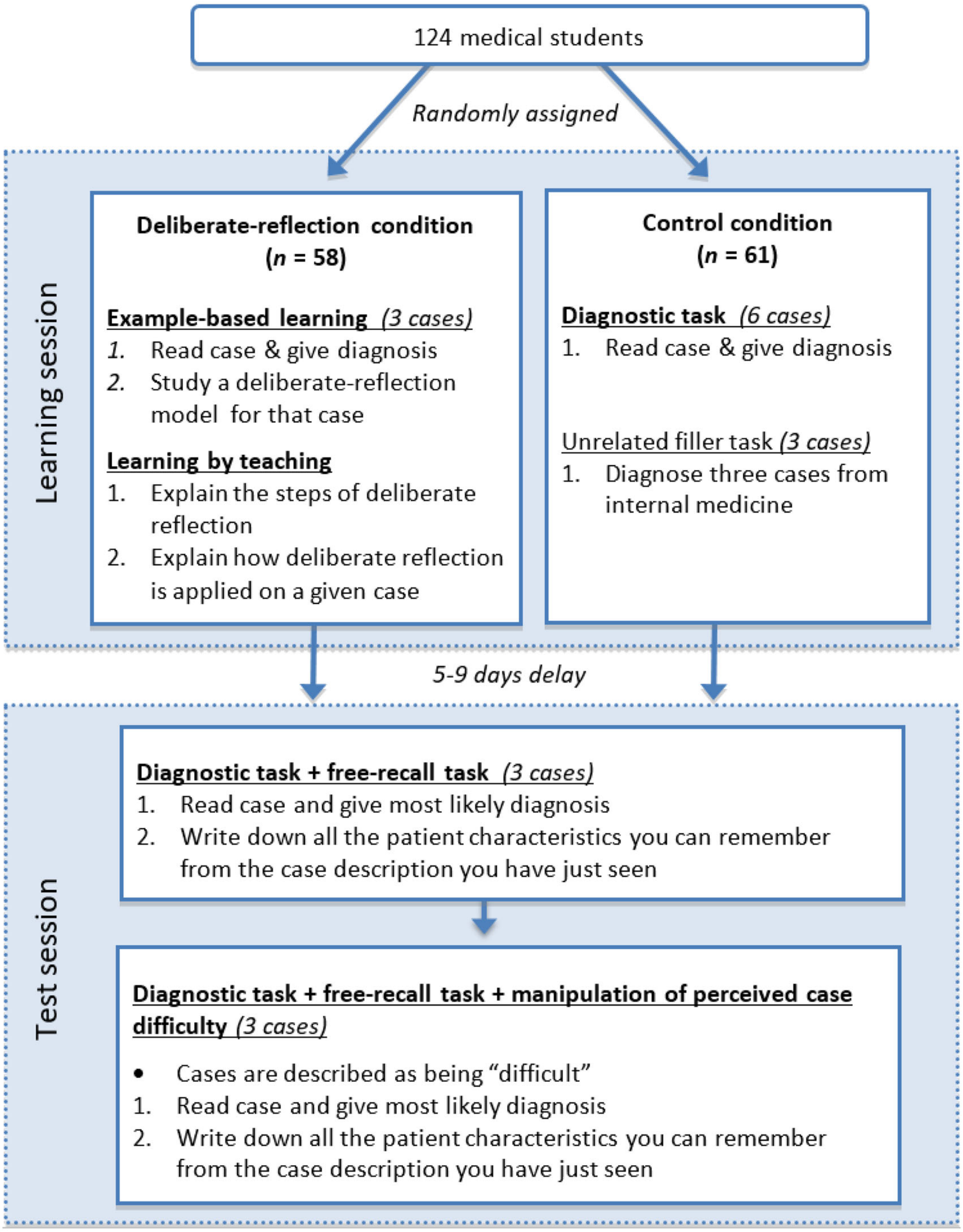
124 medical students
Randomly assigned
Learning session
Deliberate-reflection condition (n = 58)
Example-based learning (3 cases)
1. Read case & give diagnosis
2. Study a deliberate-reflection model for that case
Learning by teaching
1. Explain the steps of deliberate reflection
2. Explain how deliberate reflection is applied on a given case
Control condition (n = 61)
Diagnostic task (6 cases)
1. Read case & give diagnosis
Unrelated filler task (3 cases)
1. Diagnose three cases from internal medicine
5-9 days delay
Test session
Diagnostic task + free-recall task (3 cases)
1. Read case and give most likely diagnosis
2. Write down all the patient characteristics you can remember from the case description you have just seen
Diagnostic task + free-recall task + manipulation of perceived case difficulty (3 cases)
• Cases are described as being "difficult"
1. Read case and give most likely diagnosis
2. Write down all the patient characteristics you can remember from the case description you have just seen

Figure 1. Overview of the study design.

Master programme, ten more of these small group sessions take place.

Material and procedure

The study was conducted on computers (either laptops or desktops) at the Erasmus Medical Centre. If participants could not attend the sessions at the institution, they were allowed to complete one or both sessions at home on their own computer. The programs were computed with Qualtrics software (Version 04.2019). We prepared two versions of each program of the learning session, and four versions of the test sessions, presenting the cases in a different order.

Twelve written cases were used in this study, excluding the filler task (Table 1). The cases resembled consultations in general practice, each describing a different patient. The six cases shown during the learning phase had been prepared and validated for previous studies (Kuhn et al. 2020). The six cases shown during the test session had been intentionally prepared by experienced general practitioners to be ambiguous cases with two equally plausible diagnoses. Each of the diagnoses had the same number of *discriminating features* in the case, i.e. characteristics or symptoms that spoke only for this diagnosis and not for the other. To validate the cases, two general practitioners who were blind to the intended diagnoses independently solved the cases and were asked to give multiple diagnoses and estimate the likelihood for these diagnoses. If they did not come to the same conclusion, or they did not think that the two main diagnoses were equally likely, they discussed and adjusted the cases until they agreed.

Table 1. Clinical conditions that were presented in the cases, excluding cases of the unrelated filler task.

Learning session	Delayed test session
Both conditions:	pneumonia/ pulmonary embolism (10 d.f.)
irritable bowel syndrome (IBS)	myocardial infarction/ stomach ulcer (10 d.f.)
chronic pancreatitis	migraine/ subarachnoidal haemorrhage (8 d.f.)
irritable bowel disease (IBD)	stomach ulcer/ cholelithiasis (10 d.f.)
Control condition only:	Gout/ cellulite (4 d.f.)
Bell's palsy	lung carcinoma/ chronic obstructive pulmonary disease (COPD) (10 d.f.)
rosacea	
multiple sclerosis	

For the cases in the test session, the number of discriminating features (d.f.) that were counted for both diagnoses together, are shown in brackets. Six cases were used for each session. During the learning session, the control condition did an additional filler task where they diagnosed three cases from internal medicine, which were unrelated to the general practice cases and showed patients with acute prostatitis, acute glomerulonephritis and deep vein thrombosis.

The physicians also discussed and agreed upon a list of the discriminating features for both diagnoses.

Two weeks before a group of students could participate in the first session of the study, a researcher visited that group during one of their educational classes, informed them about the study and asked for their participation. If a student wanted to participate but was not able to attend the study session at the Erasmus Medical Centre, the researcher collected their email address. Shortly before the session, these students who were unable to attend were sent the material. For the learning session, they received an information letter, informed consent letter and a link to the study in Qualtrics. For the test session they received the link to that part of the study in Qualtrics.

Learning session

At the beginning of the first session, all participants at the Erasmus Medical Centre received an information letter and were asked to give written informed consent. Then, they started the program on the computer. All participants individually watched a video that explained the instructions for their condition and showed an example case that was diagnosed following these instructions. After this, they started to diagnose the first case.

Deliberate-reflection condition

Participants in the deliberate-reflection condition were shown a case, asked to read it and when they had come to the most likely diagnosis for the case, move on to the next page. Here they were asked to fill in the most likely diagnosis for the case (*diagnostic task)*. On the next page, they saw the case again, together with a reflection table (example in Supplementary Figure 1) that had been prepared by an experienced general practitioner. It showed the steps of deliberate reflection applied to this case with three probable diagnoses for the case. The steps of deliberate reflection are to write down the first diagnosis and then systematically list findings from the case that speak (1) for the diagnosis, (2) against the diagnosis, (3) that were absent in the case but would be expected if the diagnosis were true, then (4) considerer an alternative diagnosis and repeat steps 1–3. Only the last step, which is the ranking of the diagnosis, was left out in the shown example. Participants were asked to study the table and pay attention to the procedure (example-based learning task), and then rank the three given diagnoses themselves. On the following two pages, they were asked to rate the mental effort they invested in solving the case (Paas 1992) and their confidence in the final diagnosis. Confidence and mental effort were rated on 9-point-Likert scales ranging from 1 (very, very little confidence/effort) to 9 (very, very much confidence/effort). Then they moved on to the next case and followed the same instructions until all three cases had been diagnosed.

After this, they started with the *explanation task* for which they were asked to record two videos and explain what they had learned in this session, addressing a fictitious peer. For recording the videos, they used their webcam and an online video recorder (www.addpipe.com) that was embedded in the program. For the first video explanation activity, participants were shown an empty table with the same format as the examples of deliberate reflection they had seen before, and were asked to explain the steps of deliberate reflection and why they help to prevent common reasoning errors. For the second video explanation activity, they were shown one of the cases they had diagnosed earlier, together with a table that showed the steps of deliberate reflection, but left out the findings from the case. Participants were asked to explain how that case was solved by applying deliberate reflection and therefore how the table could be filled in while addressing a fictitious peer who was also seeing the case and table but had not solved the case. At the end of the session, participants were asked to give some demographic information.

Control condition

Participants in the control condition did the same diagnostic task as did the deliberate-reflection condition, but for six instead of three cases to increase the time spend with the study material, including the mental-effort and confidence ratings after each case. To increase the duration of the learning session even further, in order to keep it approximately the same for both conditions, participants in the control condition did an unrelated filler task. There were asked to diagnose three more cases from internal medicine, that were unrelated to the cases used in this study. At the end of the session, participants were asked to give some demographic information.

Test session

The test session was the same for both conditions. Again, participants started with a *diagnostic task*, followed by the mental-effort and confidence ratings. Then, participants completed the *recall task* in which they were asked to write down everything they could remember from the case. Then, they moved on to the next case. After having seen the first three cases, participants were told that the following three cases would be more difficult cases, that had often been misdiagnosed by students with their level of experience. As the order of the cases was counterbalanced for the four versions of the test session, case difficulty did not actually differ.

Analysis

Test phase

Participants' answers on the recall task were scored by two research assistants. They counted how many of the previously defined discriminating features for a diagnosis had been recalled. The data of 16 participants (13%) was scored by both research assistants with *excellent* interrater reliability, ICC = .90 (Cicchetti 1994).

The participants' answers on the diagnostic task were categorized into three categories by an experienced general practitioner. Category A or B were corresponding to the two diagnoses that we had determined to be the most likely for the case. We included all participants' diagnoses that were exactly the same as the two most likely diagnoses or a related diagnosis if the related diagnosis had the same discriminating features as those that we focussed on (e.g. gastritis was put in a category with stomach ulcer). Category C contained all other diagnoses, that did not fit with one of the two diagnoses.

We then calculated how many discriminating features participants had recalled that fitted with the diagnosis they had given themselves, and how many fitted with the alternative diagnosis. All instances where a participant had given a diagnosis that did not fit with one of the two most likely diagnoses for the case (category C), were excluded from this analysis ($n = 13$). We only excluded data regarding that specific medical case, we did not entirely exclude that participant's data. For one participant, we had to exclude two cases. For all other participants, we had to exclude no more than one case. For each participant, we calculated the *proportion of recalled discriminating features* by dividing the number of discriminating features that a participant had recalled by the number of features that could have been recalled for the case (ranging from 2 to 5 features per diagnosis).

For all outcome measures (time to diagnosis, proportion of recalled discriminating features for both diagnoses, proportion of recalled discriminating features for own diagnosis, proportion of recalled discriminating features for alternative diagnosis, mental effort, confidence) we calculated the mean for all cases that had no description of difficulty and the mean for the cases described as difficult. For each outcome measure, we conducted a mixed repeated-measures ANOVA with the description of difficulty as a within-subjects factor (no description; described as difficult) and study condition as a between-subjects factor (control; deliberate reflection). We used a significance level of α = .05 for all analyses and provide ηp^2 as a measure of effect size for the analyses of variances (ANOVA), with .01, .06, .14 corresponding to small, medium and large effects (Cohen 1988). When we found a significant interaction effect, we also did a simple effects test to better understand the effect that study condition had on the cases that were or were not described as difficult. For this, we conducted a one-way ANOVA for the cases of each description of difficulty separately with study condition as a between-subjects factor. The data were analysed using IBM SPSS Statistics for Windows version 25 (IBM, New York).

Results

Means and standard deviations are shown in Table 2. The analysis of *time to diagnose* showed no main effect of condition, F (1,117) = 1.33, p = .25, $\eta_p{}^2$ = .01, no main effect of description of difficulty, F (1,117) = 2.68, p = .10, $\eta_p{}^2$ = .02, and no interaction effect, F (1,117) < 0.01, p = .97, $\eta_p{}^2$ < .01.

Figure 2 depicts the three types of mean proportion of recalled discriminating features in relation to the description of difficulty of the cases. The analysis of *proportion of recalled discriminating features for both diagnoses* showed a significant main effect of condition, F (1,117) = 4.70, p = .03, $\eta_p{}^2$ = .04, as the deliberate-reflection condition (M = .63, SD = .12) recalled a higher proportion than the control condition (M = .57, SD = .14). It showed no main effect of description of difficulty, F (1,117) = 1.62, p = .21, $\eta_p{}^2$ = .01. However, there was a significant interaction effect, F (1,117) = 4.77, p = .03, $\eta_p{}^2$ = .03, indicating that description of difficulty had different effects on the two conditions. Further tests showed, that for the cases without description of difficulty the control condition recalled significantly less discriminating features than did the deliberate-reflection condition, F (1,117) = 8.85, p < .01, $\eta_p{}^2$ = .07. For the cases that were described as difficult, both conditions recalled approximately the same number of features, F (1,117) =0.89, p = .35, $\eta_p{}^2$ < .01.

The analysis of *proportion of recalled discriminating features for own diagnosis* showed no main effect of condition, F (1,117) = 1.52, p = .22, $\eta_p{}^2$ = .01, and no main effect of description of difficulty, F (1,117) = 1.25, p = .27, $\eta_p{}^2$ = .01. However, there was a significant interaction effect, F (1,117) = 9.85, p < .01, $\eta_p{}^2$ = .08. Further tests showed, that for the cases without description of difficulty the control condition recalled significantly less discriminating features related to their own diagnosis than did the deliberate-reflection condition, F (1,117) = 8.54, p < .01, $\eta_p{}^2$ = .07. For the cases that were described as difficult, both conditions recalled about the same number of features related to their own diagnosis, F (1,117) =0.44, p = .51, $\eta_p{}^2$ < .01.

The analysis of *proportion of recalled discriminating features for alternative diagnosis* showed a significant main effect of condition, F (1,117) = 6.36, p = .01, $\eta_p{}^2$ = .05, indicating that the deliberate-reflection condition scored higher (M = .60, SD = .14) than the control condition (M = .53, SD = .18). It also showed a significant main effect of description of difficulty, F (1,117) = 6.08, p = .02, $\eta_p{}^2$ = .05, as a higher proportion of discriminating features were being recalled for cases described as difficult (M = .59, SD = .20) than for cases without description of difficulty (M = .54, SD = .20). There was no significant interaction effect, F (1,117) < 0.01, p = .96, $\eta_p{}^2$ < .01.

The analysis of *mental effort* ratings showed no main effect of condition, F (1,117) < 0.01, p = .94, $\eta_p{}^2$ < .01. However, there was a significant main effect of description of difficulty, F (1,117) = 34.03, p < .01, $\eta_p{}^2$ = .23, indicating that participants reported having invested more mental effort when diagnosing cases described as difficult (M = 5.28,

Table 2. Means and standard deviations for all outcome measures collected during the test phase.

		Cases without description		Cases described as difficult		All cases	
	N	Mean	SD	Mean	SD	Mean	SD
Time to Diagnose							
Control	61	104.53	46.85	110.77	37.63	107.65	36.66
Deliberate Reflection	58	112.06	41.63	118.63	42.32	115.34	36.23
Total	119	108.20	44.36	114.60	40.01	111.40	36.50
Proportion of Recalled Discriminating Features for Both Diagnoses							
Control	61	.55	.16	.59	.15	.57	.14
Deliberate Reflection	58	.63	.13	.62	.15	.63	.12
Total	119	.59	.15	.61	.15	.60	.14
Proportion of Recalled Discriminating Features for Own Diagnosis							
Control	61	.60	.17	.63	.18	.62	.14
Deliberate Reflection	58	.69	.16	.61	.19	.65	.15
Total	119	.64	.17	.62	.18	.63	.15
Proportion of Recalled Discriminating Features for Alternative Diagnosis							
Control	61	.50	.23	.55	.19	.53	.18
Deliberate Reflection	58	.58	.16	.63	.20	.60	.14
Total	119	.54	.20	.59	.20	.56	.16
Confidence							
Control	61	6.17	1.29	5.35	1.43	5.76	1.20
Deliberate Reflection	58	6.01	1.23	5.43	1.42	5.72	1.20
Total	119	6.09	1.26	5.39	1.42	5.74	1.19
Mental Effort							
Control	61	4.49	1.54	5.27	1.54	4.88	1.38
Deliberate Reflection	58	4.52	1.52	5.28	1.49	4.90	1.29
Total	119	4.50	1.52	5.28	1.51	4.89	1.33

Time to diagnose was measured in seconds; all proportions of recalled discriminating features range from 0–1; confidence and mental effort were scored on 9-point-Likert scales ranging from 1 (very, very little confidence/effort) to 9 (very, very much confidence/effort).

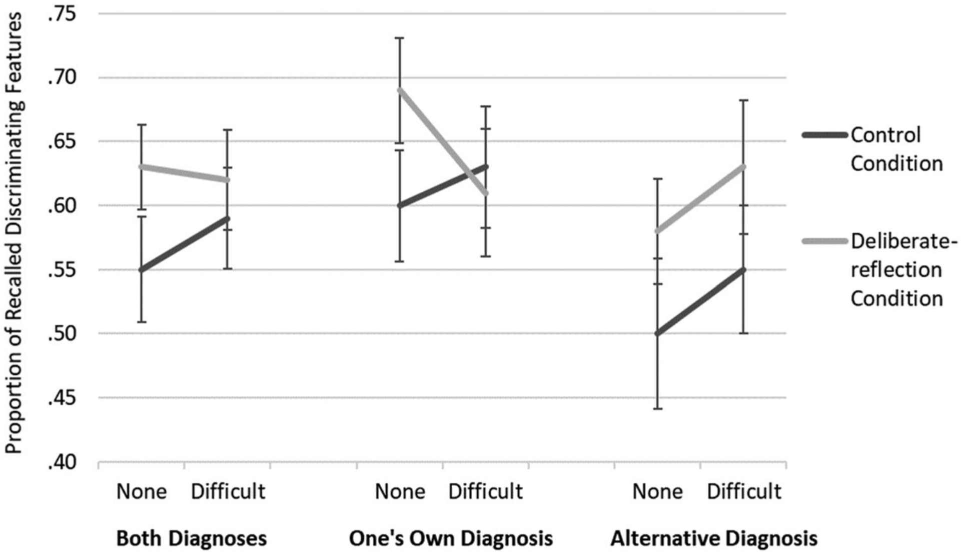

Figure 2. Mean proportion of recalled discriminating features for both diagnoses, one's own diagnosis, and the alternative diagnosis, split up by description of difficulty (none, difficult). error bars show ± 2 standard error.

$SD = 1.51$) than when diagnosing cases without description of difficulty ($M = 4.50$, $SD = 1.52$). There was no significant interaction effect, F (1,117) $= 0.01$, $p = .93$, $\eta_p{}^2 < .01$.

The analysis of *confidence* ratings showed no main effect of condition, F (1,117) $= 0.04$, $p = .85$, $\eta_p{}^2 < .01$, but there was a significant main effect of description of difficulty, F (1,117) $= 38.92$, $p < .01$, $\eta_p{}^2 = .25$, indicating that participants reported lower confidence in their diagnosis on cases described as difficult ($M = 5.39$, $SD = 1.42$) than on cases without description of difficulty ($M = 6.09$, $SD = 1.26$). There was no significant interaction effect, F (1,117) $= 1.07$, $p =$ .30, $\eta_p{}^2 = .01$.

Discussion

The current study investigated whether medical students who practised with the deliberate-reflection procedure would adopt key elements of deliberate reflection when diagnosing future cases and whether or not this would only show when they expected to diagnose difficult cases. In a learning phase, students either first studied examples of deliberate reflection and then explained the procedure to a fictitious peer on video (deliberate-reflection condition), or they diagnosed cases without deliberate reflection (control condition). In a test phase about a week later, all participants completed a diagnostic and a recall task for six ambiguous cases. Our findings show that students can indeed learn the deliberate reflection procedure *via* example-based learning and learning by teaching (cf. Kuhn et al. 2021), and more importantly, that they did in fact seem to apply it autonomously (without being triggered by a description of case difficulty) when diagnosing new cases five to nine days later.

That is, we found that students in the deliberate-reflection condition recalled more of the discriminating features of a case than did students in the control condition. This suggests that they engaged in more reflective reasoning (Mamede et al. 2007; Mamede et al. 2008), making use of key elements of deliberate reflection, especially because this effect was most pronounced for the recalled features for the alternative diagnosis that they had not given themselves (and the deliberate-reflection procedure entails considering features of alternative diagnoses). Interestingly, the application of the deliberate-reflection procedure did not seem to take more time or effort, as there were no differences between the conditions in the time needed to solve the test cases or in mental effort investment during the test phase. An explanation for this may be that while students approached the cases more reflectively, they did not apply the whole deliberate-reflection procedure.

Whilst we had expected that the effect of practising with deliberate reflection would only show on cases described as difficult, as this may trigger them to use the reflection method they had learned (Mamede et al. 2008), the opposite was true: When cases had no description of difficulty, participants in the deliberate reflection condition recalled more discriminating features overall and related to their own diagnoses than did the control condition. When they expected the cases to be difficult, however, the difference between the condition diminished and was no longer significant, as participants in the control condition also recalled more of these features. This is in line with research showing that describing a case as difficult can already induce reflective reasoning (Mamede et al. 2008; Noyer et al. 2017). This may indicate that the control condition was also able to engage in reflective reasoning and pay more attention to the details of a case, but they needed a trigger to make use of it while the deliberate-reflection condition already engaged in reflective reasoning. That the difficulty announcement served as a trigger is suggested by the fact that when cases were described as difficult, both conditions recalled more features of the alternative diagnoses than they did for cases without the description of difficulty. Moreover, the mental effort and confidence ratings confirm that manipulating the description of difficulty did indeed change the participants' perception of the case; Participants reported more mental effort investment but lower confidence for cases that they were told would be difficult cases, although the difficulty of the cases did not change.

A limitation of the study is that we do not know how long-lasting the effects of practising with deliberate reflection are. It is known that physicians' reasoning changes as they gain more experience (Schmidt and Boshuizen 1993; Schmidt and Rikers 2007). It may be, that the effect of experience reduces the effect of learning deliberate reflection. Also, the effects found during the recall task, though statistically significant, were only medium to small in size. Therefore, the results and the recommendations for educational practice should be interpreted with caution. Another limitation is, that we cannot determine whether students in the deliberate-reflection condition did apply the procedure the way they had learned it, or whether they just used elements of it.

As the purpose of this study was to see whether we could teach students deliberate reflection with a learning-by-teaching approach, we did not measure whether this would also lead to an improvement in diagnostic accuracy. The ambiguous cases used in the test phase, which had multiple possible diagnoses, were not designed to test diagnostic performance. Future studies could test whether teaching deliberate reflection to students or manipulating their perceived difficulty of a case would result in better diagnostic performance. Furthermore, it would be interesting in future research to test whether this intervention and measurement approach is only effective when used with students or also with more experienced residents and physicians, with whom we did not find an effect in previous studies (Kuhn et al. 2020, 2021).

In conclusion, the current study showed that learning deliberate reflection helped students to focus on the relevant features of a case and to avoid a tunnel vision where they only focussed on one diagnosis when they diagnosed cases one week later. Telling students that the cases would be difficult had a similar effect, but students who had learned deliberate reflection did not seem to need this information to engage in reflective reasoning. This suggests that deliberate reflection can be taught and then applied autonomously, which means that they may also be able to apply this in medical practice where they get no prompts to do so. By learning deliberate reflection as students, they may internalise this way of diagnostic reasoning, which can prevent diagnostic error in the future (Croskerry 2003; Berner and Graber 2008). Future studies could investigate whether these findings also apply to physicians with a different level of expertise and whether this would lead to an improvement in diagnostic accuracy, for example in situations when they could be misled by cognitive bias.

Acknowledgements

The authors thank all the students who participated in this study, Ronald van der Knaap and Michiel Bos for validating the recall cases, Jessica de Graaf and Jacky Hooftman for their help with the data collection and with coding and preparing the data, Inge Otto and Justine Staal for helping out with data collection, and Alexander Hoogendijk for this technical support during data collection.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by ZonMW under Grant 839130007.

Notes on contributors

Josepha Kuhn, PhD, is a former PhD candidate at the Department of General Practice and the Institute of Medical Education Research Rotterdam at the Erasmus Medical Centre, The Netherlands.

Sílvia Mamede, MD, PhD, is an associate professor at the Institute of Medical Education Research Rotterdam, Erasmus Medical Centre, The Netherlands and at the Department of Psychology, Education and Child Studies, Erasmus University Rotterdam, The Netherlands .

Pieter van den Berg, MD, PhD, is a general practitioner and former research coordinator at the Department of General Practice, Erasmus Medical Centre, The Netherlands.

Laura Zwaan, PhD, is an assistant professor at the Institute of Medical Education Research Rotterdam, Erasmus Medical Centre, The Netherlands.

Gijs Elshout, MD, PhD, is a general practitioner and an assistant professor at the Department of General Practice, Erasmus Medical Centre, The Netherlands.

Patrick Bindels, MD, PhD, is professor and the Head of the Department of General Practice, Erasmus Medical Centre, The Netherlands.

Tamara van Gog, PhD, is professor at the Department of Education, Utrecht University, The Netherlands.

ORCID

Josepha Kuhn http://orcid.org/0000-0003-1556-2957
Gijs Elshout http://orcid.org/0000-0002-6988-0179

Data availability statement

The anonymised data set can be requested from the corresponding author.

References

Berner ES, Graber ML. 2008. Overconfidence as a cause of diagnostic error in medicine. Am J Med. 121(5 Suppl):S2–S23. doi:10.1016/j.amjmed.2008.01.001.

Cicchetti DV. 1994. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychol Assess. 6(4):284–290. doi:10.1037/1040-3590.6.4.284.

Cohen J. 1988. Statistical power analysis for the behavioral sciences. 2 ed. Hillsdale: Lawrence Erlbaum Associates.

Costa Filho GB, Moura AS, Brandão PR, Schmidt HG, Mamede S. 2019. Effects of deliberate reflection on diagnostic accuracy, confidence and diagnostic calibration in dermatology. Perspect Med Educ. 8(4):230–236. doi:10.1007/s40037-019-0522-5.

Croskerry P. 2003. The importance of cognitive errors in diagnosis and strategies to minimize them. Acad Med. 78(8):775–780. doi:10.1097/00001888-200308000-00003.

Hoogerheide V, Renkl A, Fiorella L, Paas F, Van Gog T. 2019. Enhancing Example-Based Learning: teaching on Video Increases Arousal and Improves Problem-Solving Performance. J Educ Psychol. 111(1):45–56. doi:10.1037/edu0000272.

Ibiapina C, Mamede S, Moura A, Elói-Santos S, Van Gog T. 2014. Effects of free, cued and modelled reflection on medical students' diagnostic competence. Med Educ. 48(8):796–805. doi:10.1111/medu.12435.

Kostopoulou O, Russo JE, Keenan G, Delaney BC, Douiri A. 2012. Information Distortion in Physicians' Diagnostic Judgments. Med Decis Making. 32(6):831–839. doi:10.1177/0272989X12447241.

Kuhn J, Mamede S, Van den Berg P, Zwaan L, van Peet PG, Bindels P, Van Gog T. Forthcoming 2021. Learning deliberate reflection in medical diagnosis: does Learning-by-Teaching Help? Manuscript submitted for publication.

Kuhn J, Van den Berg P, Mamede S, Zwaan L, Diemers A, Bindels P, Van Gog T. 2020. Can we teach reflective reasoning in general-practice training through example-based learning and learning by doing? Health Prof Educ. 6(4):506–515. doi:10.1016/j.hpe.2020.07.004.

Lambe KA, Reilly G, Kelly BD, Curristan S. 2016. Dual-process cognitive interventions to enhance diagnostic reasoning: a systematic review. BMJ Qual Saf. 25(10):808–820. doi:10.1136/bmjqs-2015-004417.

Mamede S, Figueiredo-Soares T, Elói Santos SM, Faria RMD, Schmidt HG, Gog T. 2019. Fostering novice students' diagnostic ability: the value of guiding deliberate reflection. Med Educ. 53(6):628–637. doi:10.1111/medu.13829.

Mamede S, Schmidt HG, Penaforte JC. 2008. Effects of reflective practice on the accuracy of medical diagnoses. Med Educ. 42(5):468–475. doi:10.1111/j.1365-2923.2008.03030.x.

Mamede S, Schmidt HG, Rikers RM, Penaforte JC, Coelho-Filho JM. 2008. Influence of perceived difficulty of cases on physicians' diagnostic reasoning. Acad Med. 83(12):1210–1216. doi:10.1097/ACM.0b013e31818c71d7.

Mamede S, Schmidt HG, Rikers RMJP, Penaforte JC, Coelho-Filho JM. 2007. Breaking down automaticity: case ambiguity and the shift to reflective approaches in clinical reasoning. Med Educ. 41(12):1185–1192. doi:10.1111/j.1365-2923.2007.02921.x.

Mamede S, Van Gog T, Moura AS, De Faria RM, Peixoto JM, Rikers RM, Schmidt HG. 2012. Reflection as a strategy to foster medical students' acquisition of diagnostic competence. Med Educ. 46(5):464–472. doi:10.1111/j.1365-2923.2012.04217.x.

Mamede S, Van Gog T, Van den Berge K, Rikers RM, Van Saase JL, Van Guldener C, Schmidt HG. 2010. Effect of availability bias and reflective reasoning on diagnostic accuracy among internalmedicine residents. JAMA. 304(11):1198–1203. doi:10.1001/jama.2010.1276.

Noyer AL, Esteves JE, Thomson OP. 2017. Influence of perceived difficulty of cases on student osteopaths' diagnostic reasoning: a cross sectional study. Chiropr Man Therap. 25:32–32. doi:10.1186/s12998-017-0161-z.

Paas F. 1992. Training strategies for attaining transfer of problem-solving skill in statistics: a cognitive-load approach. J Educ Psychol. 84(4):429–434. doi:10.1037/0022-0663.84.4.429.

Prakash S, Sladek RM, Schuwirth L. 2019. Interventions to improve diagnostic decision making: a systematic review and meta-analysis on reflective strategies. Med Teach. 41(5):517–524. doi:10.1080/0142159X.2018.1497786.

Schmidt HG, Boshuizen HPA. 1993. On Acquiring Expertise in Medicine. Educ Psychol Rev. 5(3):205–221. doi:10.1007/BF01323044.

Schmidt HG, Rikers RMJP. 2007. How expertise develops in medicine: knowledge encapsulation and illness script formation. Med Educ. 41(12):1133–1139.

Schmidt HG, Van Gog T, Schuit SC, Van den Berge K, Van Daele PL, Bueving H, Van der Zee T, Van den Broek WW, Van Saase JL, Mamede S. 2017. Do patients' disruptive behaviours influence the accuracy of a doctor's diagnosis? A randomised experiment. BMJ Qual Saf. 26(1):19–23. doi:10.1136/bmjqs-2015-004109.

Van Gog T, Rummel N. 2010. Example-based learning: integrating cognitive and social-cognitive research perspectives. Educ Psychol Rev. 22(2):155–174. doi:10.1007/s10648-010-9134-7.

Wallsten TS. 1981. Physician and medical student bias in evaluating diagnostic information. Med Decis Making. 1(2):145–164. doi:10.1177/0272989X8100100205.