Available online at www.sciencedirect.com

ScienceDirect

Health Professions Education 6 (2020) 506–515

www.elsevier.com/locate/hpe

Can We Teach Reflective Reasoning in General-Practice Training Through Example-Based Learning and Learning by Doing?

Josepha Kuhn [a,b,*], Pieter van den Berg [a], Silvia Mamede [b], Laura Zwaan [b], Agnes Diemers [c], Patrick Bindels [a], Tamara van Gog [d]

[a] *Department of General Practice, Erasmus Medical Centre, the Netherlands*
[b] *Institute of Medical Education Research Rotterdam, Erasmus Medical Centre, the Netherlands*
[c] *Department of General Practice and Wenckebach Institute for Education and Training, University Medical Centre Groningen, the Netherlands*
[d] *Department of Education, Utrecht University, the Netherlands*

Received 13 January 2020; revised 8 July 2020; accepted 14 July 2020
Available online 3 August 2020

Abstract

Purpose: Flaws in physicians' reasoning frequently result in diagnostic errors. The method of *deliberate reflection* was developed to stimulate physicians to deliberately reflect upon cases, which has shown to improve diagnostic performance in complex cases. In the current randomised controlled trial, we investigated whether deliberate reflection can be taught to general-practice residents. Additionally, we investigated whether engaging in deliberate reflection or studying deliberate-reflection models would be more effective.
Methods: The study consisted of one learning session and two test sessions. Forty-four general-practice residents were randomly assigned to one of three study conditions in the learning session: (1) control without reflecting ($n = 14$); (2) engaging in deliberate reflection ($n = 11$); or (3) studying deliberate-reflection models ($n = 19$). To assess learning, they diagnosed new cases in both a same-day test and a delayed test one week later. In the delayed test, participants were additionally asked to elaborate on their decisions. We analysed diagnostic accuracy and whether their reasoning contained key elements of deliberate reflection.
Results: We found no significant differences between the study conditions in diagnostic accuracy on the same-day test, $p = .649$, or on diagnostic accuracy, $p = .747$, and reflective reasoning, $p = .647$, on the delayed test.
Discussion: Against expectations, deliberate reflection did not increase future reflective reasoning. Future studies are needed to investigate whether residents either did not sufficiently learn the procedure, did not adopt it when diagnosing cases without instructions to reflect, or whether the reflective-reasoning process as itself cannot be taught.
© 2020 King Saud bin Abdulaziz University for Health Sciences. Production and Hosting by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Keywords: Reflective reasoning; General practice; Diagnostic error; Instructional design

1. Introduction

Sound clinical reasoning is a crucial factor to ensure high diagnostic performance in general practice. There has been much discussion on how to improve

* Corresponding author. Department of General Practice, Erasmus MC, P.O. Box 2040, Rotterdam, CA, 3000, the Netherlands.
E-mail address: j.kuhn@erasmusmc.nl (J. Kuhn).

Peer review under responsibility of AMEEMR: the Association for Medical Education in the Eastern Mediterranean Region.

https://doi.org/10.1016/j.hpe.2020.07.004
2452-3011/© 2020 King Saud bin Abdulaziz University for Health Sciences. Production and Hosting by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

diagnostic reasoning, but an approach whose effectiveness is empirically supported is *deliberate reflection*. Deliberate reflection aims to stimulate physicians to further reflect on their first impression of a case at hand.[1] Thereby, it could correct diagnostic errors due to excessive reliance on intuitive reasoning. Intuitive reasoning is efficient most of the time and it enables experienced physicians to make good and fast decisions. However, it may also lead to errors, for example if physicians are being influenced by irrelevant contextual factors.[2,3] If a wrong initial diagnostic hypothesis has been generated, the mistake could only be corrected by further reflection on the case.[4]

Studies on deliberate reflection have found that it can counteract diagnostic errors on complex cases[1]**,** or if physicians were distracted by irrelevant patient features[5,6] or influenced by other cases they had encountered recently (i.e., availability bias).[6,7] Deliberate reflection has been investigated as a learning tool as well. Students (4th – 6th year) who followed the deliberate-reflection procedure during practice with clinical cases showed higher diagnostic accuracy when solving similar cases one week later than students who just diagnosed the cases.[8–10]

Studies have not yet shown, however, whether physicians could also learn the deliberate-reflection procedure itself. If this is possible, physicians could spontaneously apply deliberate reflection on new cases to be solved in the future, regardless of the content and without explicit instructions to reflect on them. It is reasonable to expect that the deliberate-reflection procedure can be taught by employing instructional approaches based on example study (i.e., example-based learning, or EBL). Such approaches have proven effective to teach problem-solving skills in many domains, particularly for novice learners.[11,12] In these domains, EBL proved more effective for novices than, for instance, learning by doing (LBD), i.e., practicing of the task. According to Cognitive Load Theory, the advantages of EBL over LBD derives from the reduced amount of cognitive load it would impose on the learner.[13] Relative to LBD, the guidance that an example gives a novice learner reduces the amount of ineffective cognitive load (i.e., the investment of cognitive resources to deal with aspects of the problem that do not help learning *how* to solve the problem). In EBL, instead of being focused on finding a solution to the problem, cognitive resources can be allocated to understand the steps involved in solving the problem. It can be said, therefore, that EBL would allow for replacing the eventually ineffective cognitive load involved in LBD by effective load imposed by studying just the procedure to be used to solve a new problem. In medical education, EBL has proven effective to teach medical procedures[14] and diagnostic competence.[15] It can therefore be hypothesised that EBL would be effective to teach deliberate reflection as well, if learners have never worked with it before.

In this study, we investigated whether the deliberate-reflection procedure can be learned and then be applied autonomously on future cases, and which teaching approach is most effective for residents in general-practice training. For this purpose, we conducted an experiment consisting of a learning session, and two test sessions. In the learning session, residents solved a set of cases either without reflection (control), by following the deliberate-reflection steps (learning by doing, LBD), or by studying deliberate-reflection models (EBL). We expected that residents would learn and adopt reflective reasoning the most when practicing with reflection models and that both reflection groups would score higher than the control group (EBL > LBD > Control).

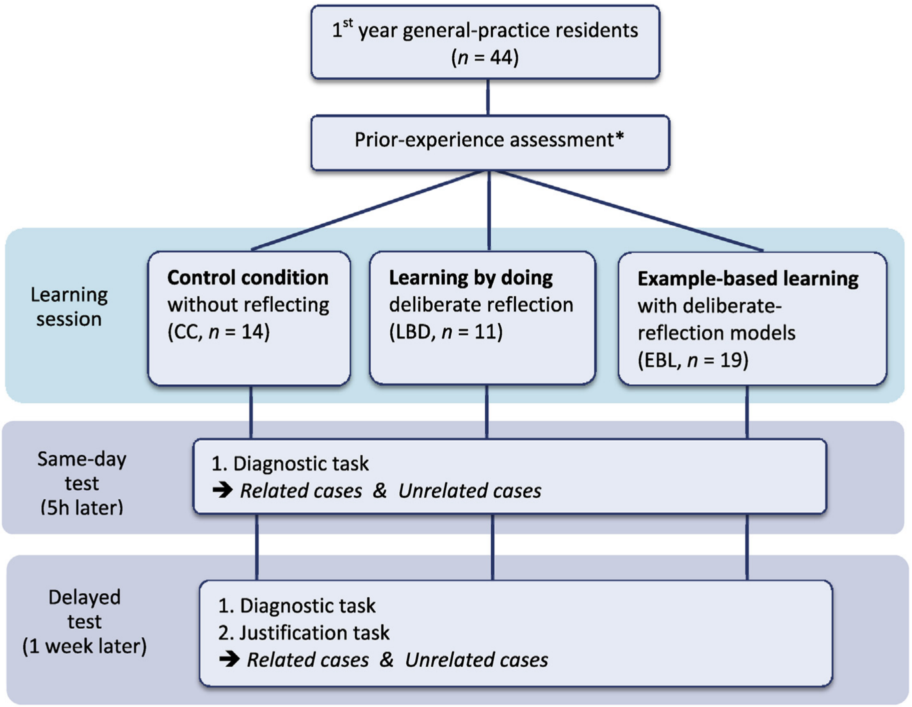
2. Method

2.1. Design

The study consisted of a prior-experience assessment and three sessions (Fig. 1): a learning session, a same-day test session, and a delayed test session. In the learning session, participants were randomly assigned to one study condition and diagnosed cases either (1) without being instructed to reflect (control); (2) by engaging in deliberate reflection (LBD); or (3) by studying deliberate-reflection models (EBL). The two test sessions were the same for all participants. The same-day test consisted of a diagnostic task, and the delayed test consisted of a diagnostic task followed by a justification task.

2.2. Participants

Eighty-one residents from the general-practice vocational training were invited to participate in the study. Participants were in the first year of a residency program at the Erasmus Medical Centre in Rotterdam or the University Medical Centre Groningen. Residents in the Netherlands have an MD degree obtained after a 6-year undergraduate training and are engaged in a three-year training program to specialize in general practice. An a priori power analysis, assuming to-be-detected effects of medium size (Cohen's $f = 0.25$)[16] at $\alpha = 0.05$, showed that a sample of 81 would be


1st year general-practice residents
(n = 44)
Prior-experience assessment*
Learning session
Control condition without reflecting (CC, n = 14)
Learning by doing deliberate reflection (LBD, n = 11)
Example-based learning with deliberate-reflection models (EBL, n = 19)
Same-day test (5h later)
1. Diagnostic task
➔ Related cases & Unrelated cases
Delayed test (1 week later)
1. Diagnostic task
2. Justification task
➔ Related cases & Unrelated cases

Note. *The prior-experience questionnaire was only filled in by 35 of the 44 participants.

Fig. 1. Illustration of the study protocol.

sufficient to have a power of 0.80. The study took place during the usual educational program, and participants did not receive compensation.

2.3. Material and procedure

All material was presented in Dutch. Thirty written cases were used in this study (Supplementary material, Table 1), each one describing a new patient. For the test sessions, 16 of the cases were related to the cases studied in the learning phase, i.e., had the same chief complaint, and eight cases were unrelated, with a completely different clinical presentation. These two types of cases were necessary to allow us to distinguish between learning the *content* of the diseases studied in the learning phase (which would show only on the related cases) and learning the de reasoning *process*, i.e., the deliberate reflection-procedure (which would show on the unrelated cases). The cases were prepared by experienced GPs, reflecting problems encountered in general practice (example in Supplementary material, Fig. 1) and validated by two different GPs. The GPs also prepared the reflection models to be used by the EBL condition (see below).

The study was presented using Qualtrics software (Version 11.2017). All participants saw the same cases during the same session. Two versions of the program were prepared for each study condition, alternating the sequence of presentation of the cases. Each session was self-paced, participants could not go back in the program, and the software automatically recorded participants' responses and time spent on each page.

2.3.1. Prior-experience questionnaire

Two weeks before the learning session, participants were asked to fill in an online questionnaire on demographics and experience in clinical practice. The questionnaire was administered in advance instead of during the study to avoid that it would influence the participants' answers during the study by priming them to diagnoses included in the questionnaire. The number and nature of new cases encountered between the prior-knowledge questionnaire and the study can be expected to be limited and without structural differences between the conditions. The questionnaire showed a list of symptoms and diagnoses, including those included in this study (Supplementary material, Table 2). For each item, participants indicated their experience on a 5-point Likert-scale ranging from 1 (I have never seen a patient with this condition, symptom, or complaint) to 5 (I have seen many patients with this condition, symptom, or complaint).

2.3.2. *Learning session*

In the learning session, participants were randomly assigned to one of three study conditions: control condition (CC), learning by doing deliberate reflection (LBD), or example-based learning with deliberate-reflection models (EBL). Before the residents arrived, we had randomly distributed papers with internet links to the different programs on the tables where the study took place. When participants arrived, we asked them to choose a table, which therefore assigned them to one of the study conditions. In advance, participants were told that the study investigated their clinical reasoning and educational methods, but they were not informed about the different conditions. Participants first watched a video with the instructions for their study condition. In the LBD and EBL condition, the video explained the steps of deliberate reflection. Thereafter, participants diagnosed six cases.

2.3.2.1. Control condition (CC). For each case, participants were requested to read a case and, as soon as they had the most likely diagnosis for the case, move on to the next page and type in the most likely diagnosis. On the next two screens, they rated their mental effort when diagnosing and their confidence in their final diagnosis, by using a 9-point-Likert-scale ranging from 1 (very low) to 9 (very high), similar to the mental-effort rating by Paas.[17] After all cases were diagnosed, the participants in the control condition did a filler task, included to ensure similar session duration across the three conditions. This filler task asked participants to diagnose four internal medicine cases, completely unrelated to the general-practice cases in this study.

2.3.2.2. Learning by doing (LBD) condition. First, participants were requested, for each case, to read the case and to give a diagnosis on the next page, just as under the control condition. Thereafter, they were asked to follow the deliberate-reflection procedure, as explained in the instruction video, to critically review the initial diagnosis.[1] Participants saw the case again with a table below. In the first row, they were asked to fill in (1) findings that support their diagnosis; (2) findings that oppose the diagnosis; (3) findings that would have been expected if the diagnosis was true but were absent; and in the next row (4) an alternative diagnosis if the diagnosis at hand turned out to be wrong. They were asked to follow the same analytical steps for this alternative diagnosis, and if possible for a third diagnosis. After this analysis, they ranked their diagnoses in order of likelihood. Finally, they rated their mental effort and confidence, and went on to the next case until all cases were diagnosed.

2.3.2.3. Example-based learning (EBL) condition. First, participants in the EBL condition read a case and gave a diagnosis, just as under the control and LBD conditions. After that, they saw the case again accompanied with a deliberate-reflection model (i.e., a filled in reflection table; example in Supplementary material, Fig. 2). The model showed the reflection table with the analysis of three plausible differential diagnoses, as used under the LBD condition. Participants were requested to study this table and, after having decided on the diagnoses' likelihood, move to the next page and fill in the ranking. Finally, they rated their mental effort and confidence, and went on to the next case until all cases were diagnosed.

2.3.3. *Same-day test session*

The same-day test was conducted three to 5 h after the learning session and was the same for all study conditions. First, participants were asked to shortly explain the diagnostic reasoning process they had applied during the first session. The purpose of this was to remind participants in the LBD and EBL condition of the deliberate-reflection steps they had learned. After this, participants diagnosed 12 new cases of which eight were *related cases* and four were *unrelated cases*. The *related cases* ($n = 16$) presented the same chief symptoms as studied cases from the learning session, either with the same or a different diagnosis. The *unrelated cases* ($n = 8$) presented novel chief symptoms and diagnoses that had not been encountered in the learning session. The procedure of the diagnostic task was the same as for the control condition in the learning session: participants read a case, went on to the next page, and gave the most likely diagnosis; they rated their mental effort and confidence, and went on to the next case until all cases were diagnosed.

2.3.4. *Delayed test session*

The delayed test was conducted seven days after the first two sessions in order to test whether a possible effect of practicing with deliberate reflection would last or would only show later. It consisted of a diagnostic task and a justification task. First, participants diagnosed, one by one, a new set of 12 cases of which eight were *related cases* and four were *unrelated cases*.

When all cases had been diagnosed, they performed a justification task that was later used to evaluate engagement in reflective reasoning. For each case, participants were shown a few sentences of the case (Supplementary material, Fig. 3), together with the diagnosis that they had given. They were asked to explain (in writing) their reasoning when making the diagnosis during the diagnostic task. Finally, participants received a written debriefing, were asked for their informed consent and thanked for their participation.

2.4. Data analysis

We used a significance level of $\alpha = .05$ and did a Bonferroni correction for the high number of tests, which led to $\alpha = .005$. As a measure of effect size, η_p^2 is provided for the analyses of variances, with .01, .06, .14 corresponding to small, medium and large effects, and r for t-tests with .10, .30, .50 as thresholds.[16]

2.4.1. Prior experience

For all chief symptoms, and for all correct diagnoses of the cases in this study, we computed the mean prior-experience ratings. On these two measures we conducted one-way analyses of variance (ANOVA) with study condition (LBD, EBL, control) as a between-subjects factor, to check for initial differences between the groups.

2.4.2. Same-day and delayed test

The accuracy of the diagnoses provided by participants was scored as either 1 (correct), 0.5 (partially correct), or 0 (incorrect). An answer was considered correct if the main component of the diagnosis appeared in it. An answer was partially correct when it contained one of the constituent elements of the diagnosis, but the core diagnosis was not cited. Incorrect answers did not cite the core diagnosis and none of its constituent elements. Each answer was scored by two general practitioners and discrepancies were resolved by discussion. The inter-rater reliability of the raters' initial scores was excellent, $ICC = .96$,.[18] The time that participants spend on a case (*time to diagnose*) was retrieved in seconds. Time to diagnose was used as an indirect measure of reflection, assuming that engaging in reflective reasoning takes more time than intuitive reasoning.

We computed the participants' mean scores on diagnostic accuracy, mental effort, confidence, and time to diagnose, separated by type of cases (related, unrelated). To test an effect of study condition on these measures, each measure was analysed by a mixed ANOVA with pairwise comparisons using Bonferroni corrections. Type of case was used as a within-subjects factor, and study condition (LBD, EBL, control) as a between-subjects factor.

The answers of the justification task were analysed for key elements of deliberate reflection. For this purpose, we counted the numbers of idea units[19,20] that could be categorised according to the deliberate-reflection steps 1−4 (example in Supplementary material, Fig. 4). An idea unit is the smallest meaningful idea that can be identified in a fragment of text. We reconstructed deliberate-reflection tables from the residents' answers, and as a result, an idea unit could be counted multiple times if it was associated with multiple diagnoses. For example, if a resident argued that a symptom speaks against two diagnoses, that symptom was counted twice. Two researchers, who were blind to the study condition, counted and categorised the idea units for 6 of the 44 participants, without judging the correctness of the medical content. The inter-rater reliability was calculated for the number of idea units per column of the reflection table and was ranging from excellent to fair[18] (left to right: $ICC = .93$, $ICC = .80$, $ICC = .85$, $ICC = .50$). Therefore, one researcher rated the complete data set.

We calculated two outcome measures about the count of idea units. As a first measure, we analysed the *number of all idea units* to see how many idea units participants generated in general. A crucial element of deliberate reflection is that participants are asked to not only consider information that supports a diagnosis at hand, but to consider contradictory arguments and alternative diagnoses also.[21] Therefore, as a second measure, we analysed the number of contradiction units in the participants reasoning to measure adoption of the deliberate-reflection procedure. *Contradiction units* were idea units counted at step 2, 3, and 4 of the deliberate-reflection procedure. For the statistical analysis, the *proportion of contradiction units* was calculated to see how many contradiction units were given relative to all idea units given by the participant. The proportions adjust for possible differences between cases in the total number of idea units that participants reported.

For the analysis, we computed the participants' *mean number of all idea units* as well as the *mean proportion of contradiction units,* separated by the two types of cases. Mixed ANOVAs with pairwise comparisons were conducted on each outcome measure with type of cases (related, unrelated) as within-

Table 1
Statistical outcomes of several ANOVA performed on the outcome data.

	Study Condition	Type of Case	Study Condition * Type of Case
Prior knowledge			
Chief symptoms	F(2, 32) = 1.03, $\eta_p^2 = .06$		
Diagnoses	$F(2, 32) = .83$, $\eta_p^2 = .05$		
Same-day test			
Diagnostic accuracy	$F(2, 41) = .44$, $p = .649$, $\eta_p^2 = .02$	$F(1, 41) = 37.34$, $p < .001$, $\eta_p^2 = .48$	$F(2, 41) = 2.07$, $p = .139$, $\eta_p^2 = .09$
Time to diagnose	$F(2, 41) = .70$, $p = .503$, $\eta_p^2 = .03$	$F(1, 41) = .18$, $p = .675$, $\eta_p^2 < .01$	$F(2, 41) = .05$, $p = .954$, $\eta_p^2 < .01$
Mental effort	$F(2, 41) = 2.64$, $p = .083$, $\eta_p^2 = .11$	$F(1, 41) = .54$, $p = .468$, $\eta_p^2 = .01$	$F(2, 41) = 2.46$, $p = .098$, $\eta_p^2 = .11$
Confidence	$F(2, 41) = 2.05$, $p = .141$, $\eta_p^2 = .09$	$F(1, 41) = .03$, $p = .870$, $\eta_p^2 < .01$	$F(2, 41) = 2.45$, $p = .099$, $\eta_p^2 = .11$
Delayed test			
Diagnostic accuracy	$F(2, 41) = .29$, $p = .747$, $\eta_p^2 = .01$	$F(1, 41) < .01$, $p = .996$, $\eta_p^2 = .00$	$F(2, 41) = 1.86$, $p = .169$, $\eta_p^2 = .08$
Time to diagnose	$F(2, 41) = 1.46$, $p = .244$, $\eta_p^2 = .07$	$F(1, 41) = 37.65$, $p < .001$, $\eta_p^2 = .48$	$F(2, 41) = .95$, $p = .393$, $\eta_p^2 = .05$
Mental effort	$F(2, 41) = 4.01$, $p = .026$, $\eta_p^2 = .16$	$F(1, 41) = .80$, $p = .378$, $\eta_p^2 = .02$	$F(2, 41) = 1.22$, $p = .305$, $\eta_p^2 = .06$
Confidence	$F(2, 41) = 2.00$, $p = .148$, $\eta_p^2 = .09$	$F(1, 41) = .24$, $p = .622$, $\eta_p^2 = .01$	$F(2, 41) = .62$, $p = .544$, $\eta_p^2 = .03$
Proportion of contradiction units	$F(2, 41) = .44$, $p = .647$, $\eta_p^2 = .02$	$F(1, 41) = 7.01$, $p = .011$, $\eta_p^2 = .147$	$F(2, 41) = .55$, $p = .624$, $\eta_p^2 = .02$
Number of all idea units	$F(2, 41) = 2.33$, $p = .110$, $\eta_p^2 = .10$	$F(1, 41) = 1.59$, $p = .214$, $\eta_p^2 = .04$	$F(2, 41) = .15$, $p = .855$, $\eta_p^2 = .08$

subjects factors, and study condition (EBL, LBD, control) as a between-subjects factor.

3. Results

3.1. Participants

Fifty-seven residents participated in the learning session (CC: $n = 19$; LBD: $n = 16$; EBL: $n = 22$) and 44 of them also completed both tests (35 female; age $M = 30.16$, $SD = 5.04$; Appendix A). Unfortunately, we had difficulties recruiting participants and as a consequence did not reach the sample size estimated by the prior power analysis. The 13 participants who did not attend the test sessions were excluded from the study, which led to unequal sample sizes of the study conditions (CC: $n = 14$; LBD: $n = 11$; EBL: $n = 19$). The study sample consisted of 14 residents from the Erasmus Medical Centre in Rotterdam and 30 residents from the University Medical Centre Groningen. Because the study sessions were considered part of the regular training, the residents could join any of the sessions. For that reason, 15 participants came to the test sessions while not having participated in the learning session and therefore were excluded from the data analysis.

3.2. Prior experience

The response rate on the prior-experience questionnaire was 79.54% (means in Appendix A). There was no difference in prior experience with the chief symptoms between the three study conditions, $p = .367$, or with the medical conditions, p = .447 (Table 1) and the three groups had similar practical/working experience in medical practice (Supplementary material, Table 3).

3.3. Same-day test

Means and standard deviations are shown in Appendix B. The ANOVA on diagnostic accuracy showed no main effect of study condition, $p = .649$, but participants performed better on related cases than on unrelated cases, $p < .001$, without a significant interaction effect, p = .139 (Table 1). The analysis on time to diagnose showed no main effect of study condition, $p = .503$, no main effect of type of case, $p = .675$, and no significant interaction, $p = .954$. The analysis on the mental effort ratings showed no main effect of study condition, $p = .083$, no main effect of type of case, $p = .468$, and no significant interaction, *p* = .098. The analysis on the confidence ratings showed no main effect of study condition, *p* = .141, no main effect of type of case, *p* = .870, and no significant interaction, *p* = .099.

3.4. Delayed test

3.4.1. Diagnostic task

Means and standard deviations are shown in Appendix C. The analysis of diagnostic accuracy showed no main effect of study condition, *p* = .747, no main effect of type of case, *p* = .996, and no significant interaction, p = .169 (Table 1). The analysis on time to diagnose showed no main effect of study condition, *p* = .244, but participants spend more time diagnosing related cases than unrelated cases, *p* < .001, without significant interaction effect, *p* = .393. The analysis of the mental effort ratings showed no main effect of study condition, *p* = .026, no main effect of type of case, *p* = .378, and no significant interaction, *p* = .305. The analysis on the confidence ratings showed no main effect of study condition, *p* = .148, no main effect of type of case, *p* = .622, and no significant interaction, *p* = .544.

3.4.2. Justification task

When analysing the data, we noticed that the elaborateness of the explanations differed much between participants. Some residents just listed a couple of findings without further explanation, which were the main findings supporting their diagnosis. Others described different diagnoses they had considered at the time of diagnosing, and which arguments influenced their estimation of likelihood. Furthermore, it was often stated that their first concern was to exclude severe diseases (e.g., cancer) before finding the most likely diagnosis. The analysis on the number of all idea units showed no main effect of study condition, *p* = .110, no main effect of type of case, *p* = .214, and no significant interaction, *p* = .855 (Table 1). The analysis on mean proportion of contradiction units showed no main effect of study condition, *p* = .647, no main effect of type of case, *p* = .011, and no significant interaction, *p* = .624.

4. Discussion

To our knowledge, this study is the first to investigate whether general-practice residents can learn the *deliberate-reflection* procedure and then adopt it autonomously when diagnosing future cases. However, our study did not show that practicing with deliberate reflection increased residents' reflective reasoning or improved their diagnostic performance, when compared to a control condition. We assumed that engaging in reflective reasoning is reflected by more time to diagnose, and more idea units and a higher proportion of contradiction units on the justification task. More reflection could then lead to higher diagnostic accuracy, when cases are difficult. Contrary to our hypotheses, the three study conditions (Control, LBD, EBL) did not differ on any of these main outcome measures. Below we will discuss why we think that we did not find LBD and EBL to be effective methods for residents to learn deliberate reflection in order to improve their reflective-reasoning skills.

The diagnostic accuracy measures show that performance was not at ceiling level and could have been improved if the residents had engaged in reflection. Therefore, there may be three possible explanations why our hypotheses were not confirmed. A first explanation is that the residents did not learn the deliberate-reflection procedure sufficiently during the learning session. It could be that one learning session was insufficient to learn the procedure, even though studies showed that it is possible to learn reasoning procedures from just one session.[22] It is also possible that they focussed more on the content of the cases rather than on the reflection procedure. A different instructional approach than EBL or LBD may be more effective to teach deliberate reflection.

A second explanation is that, even though residents learned the deliberate-reflection procedure, they did not apply it when diagnosing cases in the test sessions. One reason for that might be that residents are already too experienced with diagnosing cases, which led them to have already acquired a diagnostic reasoning approach that they routinely adopt when solving clinical problems. Consequently, the learning session may not have been sufficient to change their usual practice. Therefore, residents' experience with the task, although not with the procedure to learn, could explain why the often found benefit of EBL for teaching problem-solving skills to novices,[11,12] did not apply to them. In studies where residents' diagnostic accuracy was improved by deliberate reflection,[1,5,6] they were directly instructed to apply the procedure while solving clinical cases. It was not tested whether participants had learned the deliberate reflection procedure and would apply it by themselves on future cases. Therefore their experience with a particular reasoning approach would not have played the same role as in the current study.

A third explanation may be that the reflective-reasoning process as itself cannot be taught, because which mode a physician would engage in is determined by the interplay between the physician and the perceived case difficulty, and is unconsciously determined. This is in line with the finding that interventions focusing on the reasoning process itself are often not effective to improve diagnostic accuracy.[23,24] Content specific interventions, on the other hand, which improve or activate physicians' knowledge, often are effective. Deliberate reflection may then be a useful educational tool to improve knowledge, as has been found in earlier studies,[8–10] but not as a reasoning strategy that is applied in practice.

Another finding of our study was that on the same-day test, all study conditions scored higher on diagnostic accuracy for cases that were related to the studied cases than for those cases not related. One explanation is, that the difficulty of these case was different. However, it could also be that participants had gained knowledge of the cases' content or recognised similarities with the studies cases, which were then forgotten in the delayed test, where this finding did not reoccur.

There are several limitations of the study. First, the sample size was small which means that the results can only serve as an indication, and the prior-knowledge questionnaire we used to rule out confounders was not filled in by all participants. Second, residents practiced the reasoning approach in a single, short session and then worked in general practice for a week before they did the delayed test. Therefore, the effect of the learning session may have limited effect on their diagnostic reasoning strategy. Third, the justification task is a post hoc explanation of how residents reasoned when diagnosing a case. It might be that this task does not sufficiently reflect the actual reasoning process but rather a rationale built subsequently. Last, the same-day test could have served as another opportunity to practice with the cases for all study conditions, including the control condition (see testing effect).[25] This could have influenced the diagnostic performance for similar cases on the delayed test.

From the findings of our study, we conclude that residents may already have considerable experience in diagnosing cases, making it more difficult to influence how they reason. Therefore, it might be more effective to teach deliberate reflection early on in their education, when students start learning how to diagnose, or with a different instructional approach. It may also be, that it is not possible to learn reflective reasoning and apply it to new cases. Practicing with deliberate reflection could have content specific benefits only and be effective for diagnosing future similar cases. Finally, future studies should measure the residents' reasoning at the time that they are solving a case, as this could be a better representation of their reasoning than our justification task.

Ethical approval

The present study does not meet the criteria for medical research (Centrale Commissie Mensgebonden Onderzoek, https://www.ccmo.nl/onderzoekers/wet-en-regelgeving-voor-medisch-wetenschappelijk-onderzoek/uw-onderzoek-wmo-plichtig-of-niet; Accessed 25-10-219), which excludes this study for ethical approval. Participation in the study was anonymous and voluntary. The individual participants are not traceable in the presented data. All procedures performed were in accordance with the American Psychological Association's (APA) Ethical Principles of Psychologists and Code of Conduct.

Funding

This research was funded by ZonMW, The Netherlands [839130007].

Other disclosure

None.

Declaration of competing interest

The authors Josepha Kuhn, Pieter van den Berg, Silvia Mamede, Laura Zwaan, Agnes Diemers, Patrick Bindels, and Tamara van Gog declare that they have no conflict of interest.

Acknowledgment

The authors would like to thank all residents who were willing to participate in our study, Christine Postma, Justine Staal, and Jacky Hooftman for their help with the data collection and coding the data, Kees in 't Veld, Herman Bueving, Marloes Jacobs, Robert Zegers, Arjen Muller, Boris Schouten, Michael Kevenaar, Roos Balvert, Yvonne van Trier, Hilde van Meer, and Charlotte van Sassen for writing and evaluating the cases and the deliberate-reflection models, and Yvonne van Trier, Boris Schouten, Robert Zegers, Charlotte van Sassen, and Michiel Bos for scoring the data.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.hpe.2020.07.004.

Appendixes.

Appendix A: Age and gender of the participants, and prior-experience rating of the diagnoses and chief symptoms presented in this study

		All cases	
	N	*Mean*	*SD*
Age			
Control	14 (14 female)	29.79	6.19
EBL	19 (14 female)	29.52	3.99
LBD	11 (7 female)	31.73	5.22
Total	44 (35 female)	30.16	5.04
Prior-experience Diagnoses			
Control	12	2.52	.45
EBL	12	2.72	.52
LBD	11	2.39	.45
Total	35	2.58	.47
Prior-experience Chief complaints			
Control	12	3.03	.54
EBL	12	3.35	.58
LBD	11	3.02	.66
Total	35	3.12	.59

Note. Participants indicated their experience on a 5-point Likert-scale ranging from 1 (I have never seen a patient with this condition, symptom, or complaint) to 5 (I have seen many patients with this condition, symptom, or complaint).

Appendix B: All outcome measures of the diagnostic task collected during the same-day test

		Related cases		Unrelated cases		All cases	
	N	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*
Diagnostic Accuracy							
Control	14	.44	.15	.32	.12	.40	.09
EBL	19	.54	.21	.29	.24	.46	.19
LBD	11	.52	.17	.22	.18	.42	.14
Total	44	.50	.18	.28	.19	.43	.15
Time to Diagnose							
Control	14	109.23	30.26	103.94	33.98	107.47	27.54
EBL	19	120.57	51.48	117.92	44.12	119.69	42.01
LBD	11	108.65	25.28	108.45	29.62	108.59	25.03
Total	44	113.98	39.61	111.11	37.51	113.03	33.89
Mental Effort							
Control	14	5.46	1.10	5.18	1.06	5.37	.99
EBL	19	4.76	1.32	5.14	1.31	4.89	1.24
LBD	11	4.22	.95	4.41	1.11	4.28	.97
Total	44	4.85	1.24	4.97	1.20	4.89	1.15
Confidence							
Control	14	4.46	1.13	4.93	1.29	4.61	1.07
EBL	19	5.59	.94	5.20	1.19	5.46	.88
LBD	11	5.13	.88	4.95	1.33	5.07	.90
Total	44	5.11	1.08	5.05	1.23	5.09	1.00

Note. Diagnostic accuracy was scored as 0 (incorrect), 0.5 (partially correct), or 1 point (correct). Time to diagnose was measured in seconds. Mental Effort and Confidence were rated on a 9-point Likert-scale ranging from 1 (very low) to 9 (very high).

Appendix C: All outcome measures of the diagnostic task and the justification task collected during the delayed test

		Related cases		Unrelated cases		All cases	
	N	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*
Diagnostic accuracy							
Control	14	.68	.18	.57	.32	.65	.16
EBL	19	.64	.19	.63	.26	.63	.17
LBD	11	.53	.14	.65	.18	.57	.10
Total	44	.62	.18	.61	.26	.62	.16
Time to diagnose							
Control	14	112.22	34.20	92.16	29.01	105.53	31.27
EBL	19	128.36	38.14	109.43	24.37	122.05	31.26
LBD	11	122.46	31.04	91.21	29.58	112.04	28.78
Total	44	121.75	35.15	99.38	28.02	114.29	30.82
Mental effort							
Control	14	5.03	1.30	4.75	1.00	4.93	1.13
EBL	19	4.71	1.07	4.89	1.22	4.77	1.03
LBD	11	4.01	.60	3.68	1.37	3.90	.71
Total	44	4.64	1.11	4.55	1.28	4.61	1.06
Confidence							
Control	14	4.92	1.25	5.02	1.08	4.95	1.09
EBL	19	5.61	.86	5.47	.90	5.56	.79
LBD	11	5.43	.86	5.68	1.11	5.52	.81
Total	44	5.34	1.02	5.38	1.02	5.36	.92
Proportion of contradiction units							
Control	14	.12	.11	.15	.09	.13	.10
EBL	19	.12	.13	.19	.15	.14	.12
LBD	11	.09	.12	.14	.15	.11	.11
Total	44	.11	.12	.16	.13	.13	.11
Number of all idea units							
Control	14	5.79	1.38	5.43	1.58	5.67	1.31
EBL	19	5.79	1.33	5.63	1.49	5.74	1.30
LBD	11	4.78	1.21	4.64	1.39	4.73	1.16
Total	44	5.54	1.36	5.32	1.52	5.47	1.31

Note. Diagnostic accuracy was scored as 0 (incorrect), 0.5 (partially correct), or 1 point (correct). Time to diagnose was measured in seconds. Mental Effort and Confidence were rated on a 9-point Likert-scale ranging from 1 (very low) to 9 (very high).

References

1. Mamede S, Schmidt HG, Penaforte JC. Effects of reflective practice on the accuracy of medical diagnoses. *Med Educ*. 2008;42(5):468–475.
2. Mamede S, Van Gog T, Schuit SC, Van den Berge K, Van Daele PL, Bueving H, et al. Why patients' disruptive behaviours impair diagnostic reasoning: a randomised experiment. *BMJ Qual Saf*. 2017;26(1):13–18.

3.. Mamede S, Van Gog T, Van den Berge K, Van Saase JLCM, Schmidt HG. Why do doctors make mistakes? A study of the role of salient distracting clinical features. *Acad Med*. 2014;89(1):114–120.

4. Hess BJ, Lipner RS, Thompson V, Holmboe ES, Graber ML. Blink or think: can further reflection improve initial diagnostic impressions? *Acad Med*. 2015;90(1):112–118.

5.. Schmidt HG, Van Gog T, Schuit SCE, Van Den Berge K, Van Daele PLA, Bueving H, et al. Do patients' disruptive behaviours influence the accuracy of a doctor's diagnosis? A randomised experiment. *BMJ Qual Saf*. 2017;26(1):19–23.

6.. Mamede S, Splinter TA, Van Gog T, Rikers RM, Schmidt HG. Exploring the role of salient distracting clinical features in the emergence of diagnostic errors and the mechanisms through which reflection counteracts mistakes. *BMJ Qual Saf*. 2012;21(4):295–300.

7.. Mamede S, Van Gog T, Van den Berge K, Rikers RM, Van Saase JL, Van Guldener C, et al. Effect of availability bias and reflective reasoning on diagnostic accuracy among internal medicine residents. *J Am Med Assoc*. 2010;304(11):1198–1203.

8.. Mamede S, Van Gog T, Moura AS, De Faria RMD, Peixoto JM, 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;89:121–127.

9.. Mamede S, Van Gog T, Moura AS, De Faria RM, Peixoto JM, Rikers RM, et al. Reflection as a strategy to foster medical students' acquisition of diagnostic competence. *Med Educ*. 2012;46(5):464–472.

10.. Ibiapina C, Mamede S, Moura A, Elói-Santos S, Van Gog T. Effects of free, cued and modelled reflection on medical students' diagnostic competence. *Med Educ*. 2014;48:796–805.

11.. Van Gog T, Rummel N. Example-based learning: integrating cognitive and social-cognitive research perspectives. *Educ Psychol Rev*. 2010;22(2):155–174.

12. Atkinson RK, Derry SJ, Renkl A, Wortham D. Learning from examples: instructional Principles from the worked examples research. *Rev Educ Res*. 2000;70(2):181–214.
13. Sweller J. Cognitive load during problem solving: effects on learning. *Cognit Sci*. 1988;12(2):257–285.

14.. Bjerrum AS, Hilberg O, Van Gog T, Charles P, Eika B. Effects of modelling examples in complex procedural skills training: a randomised study. *Med Educ*. 2013;47(9):888–898.

15. Stark R, Kopp V, Fischer MR. Case-based learning with worked examples in complex domains: two experimental studies in undergraduate medical education. *Learn InStruct*. 2011;21(1):22–33.
16. Cohen J. *Statistical power analysis for the behavioral sciences*. 2 ed. Hillsdale: Lawrence Erlbaum Associates; 1988.
17. Paas FGWC. Training strategies for attaining transfer of problem-solving skill in statistics: a cognitive-load approach. *J Educ Psychol*. 1992;84(4):429–434.
18. Cicchetti DV. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychol Assess*. 1994;6(4):284–290.
19. Meyer BJF. *The organization of prose and its effects on memory*. Amsterdam: North-Holland Publishing Co. New York: American Elsevier Publishing Co.; 1975.
20. Schiefele U, Krapp A. Topic interest and free recall of expository text. *Learn Indiv Differ*. 1996;8(2):141–160.
21. Mamede S, Schmidt HG. Reflection in diagnostic reasoning: what really matters? *Acad Med*. 2014;89(7):959–960.

22.. Hoogerheide V, Loyens SMM, Van Gog T. Effects of creating video-based modeling examples on learning and transfer. *Learn InStruct*. 2014;33:108–119.

23. Norman GR, Monteiro SD, Sherbino J, Ilgen JS, Schmidt HG, Mamede S. The causes of errors in clinical reasoning: cognitive biases, knowledge deficits, and dual process thinking. *Acad Med*. 2017;92(1):23–30.
24. Schmidt HG, Mamede Sl. How to improve the teaching of clinical reasoning: a narrative review and a proposal. *Med Educ*. 2015;49(10):961–973.
25. Roediger HL, Karpicke JD. The power of testing memory: basic research and implications for educational practice. *Perspect Psychol Sci*. 2006;1(3):181–210.

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

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

Silvia Mamede is an associate professor at the Institute of Medical Education Research Rotterdam, Erasmus Medical Centre, The Netherlands.

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

Agnes Diemers is a senior researcher at the Department of Development & Research of Medical Education and the Institute of General Practice, and works as the Head of Faculty Development at the Wenckebach Institute for Education and Training, University Medical Centre Groningen, The Netherlands.

Patrick Bindels the Head of the Department of General Practice, Erasmus Medical Centre, The Netherlands.

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