Fostering novice students' diagnostic ability: the value of guiding deliberate reflection

Sílvia Mamede,[1,2] Taciana Figueiredo-Soares,[3] Silvana M Elói Santos,[3] Rosa M D de Faria,[3] Henk G Schmidt[1,2] & Tamara van Gog[4]

BACKGROUND Deliberate reflection when practising the diagnosis of clinical cases has been shown to develop medical students' diagnostic competence. Adding guidance by cueing reflection or providing modelling of reflection increased the benefits of reflection for advanced (Years 5–6) students. The present study investigated whether we could replicate and extend these findings by comparing the effects of free, cued and modelled reflection on novice students' diagnostic competence.

METHODS A total of 80 third-year medical students participated in a two-phase experiment. In the learning phase, students diagnosed nine clinical cases under one of three conditions: free reflection; cued reflection, and modelled reflection. Two weeks later, all students diagnosed four new examples of the diseases studied in the learning phase and four cases of non-studied related diseases ('adjacent diseases'). The main outcome measurements were diagnostic accuracy scores (range 0–1) on studied and adjacent diseases.

RESULTS For studied diseases, there was a significant effect of experimental condition on diagnostic accuracy ($p < 0.02$), with the cued-reflection group (mean = 0.58, standard deviation [SD] = 0.23) performing significantly better than the free-reflection group (mean = 0.41, SD = 0.20; $p < 0.02$). The cued-reflection and modelled-reflection groups (mean = 0.54, SD = 0.22) did not differ in diagnostic accuracy ($p > 0.05$), nor did the modelled-reflection group perform better than the free-reflection group ($p > 0.05$). For adjacent diseases, the three groups scored extremely low, without significant differences in performance ($p > 0.05$). Cued reflection and free reflection were rated as requiring similar effort ($p > 0.05$) and both were more demanding than studying examples of reflection (both $p < 0.001$) in the learning phase.

CONCLUSIONS Simply cueing novice students' reflection to focus it on relevant diseases was sufficient to increase diagnostic performance relative to reflection without any guidance. Cued reflection and studying examples of reflection appear to be equally useful approaches for teaching clinical diagnosis to novice students. Students found studying examples of reflection required less effort but cued reflection will certainly demand much less investment from teachers.

Medical Education 2019: 53: 628–637
doi: 10.1111/medu.13829

[1]Institute of Medical Education Research Rotterdam, Erasmus MC, University Medical Center Rotterdam, Rotterdam, the Netherlands
[2]Department of Psychology, Erasmus University, Rotterdam, the Netherlands
[3]Department of Propedeutics, Faculty of Medicine, Federal University of Minas Gerais, Belo Horizonte, Brazil
[4]Department of Education, Utrecht University, Utrecht, the Netherlands

Correspondence: Sílvia Mamede, Institute of Medical Education Research Rotterdam, Erasmus MC, University Medical Center Rotterdam, Erasmus Medical Center – Ae-242, Wytemaweg 80, 3015 CN Rotterdam, the Netherlands. Tel: 00 31 10 703 8270; E-mail: s.mamede@erasmusmc.nl

INTRODUCTION

The quality of patient care depends critically on physicians' ability to make sound clinical judgements. Not surprisingly, medical education places much emphasis on the teaching of clinical reasoning.[1] However, there is scarce empirical evidence on which teaching approach works better under which circumstances and for whom.[1,2] This leaves clinical teachers without enough direction in choosing, for instance, the instructional approach to be used when students' practice with clinical cases. Because exposure to a variety of clinical problems is crucial for the development of clinical reasoning,[3] practice with cases is the core element of the clinical reasoning courses now offered by many schools. The approaches employed in these courses vary widely, but empirical research on their effectiveness is scarce.[4] One of the few approaches supported by experimental evidence is deliberate reflection. Developed and studied initially in the context of physicians' diagnostic performance,[5,6] it was subsequently investigated as a learning tool, and shown to have a large positive effect on medical students' diagnostic competence.[7,8] The purpose of the present study was to investigate whether this positive effect of deliberate reflection can be enhanced by providing students with additional guidance as they reflect upon clinical cases.

The deliberate reflection procedure consists of comparing alternative diagnoses for a case and reflecting upon evidence from the case that is (and is not) in line with the diagnoses.[5] In two experiments, medical students who followed the deliberate reflection procedure during practice with clinical cases made better diagnoses of new cases of the same (and related) diseases in a delayed test than students who practised by providing immediate or differential diagnoses.[7,8] The positive effect of deliberate reflection was attributed to a refinement of the nascent illness scripts stored in the students' memory. An illness script is a mental representation of a disease, containing knowledge of its causal mechanisms, clinical presentations and the conditions under which it occurs.[9,10] When activated by patients' cues, illness scripts bring diagnostic hypotheses to mind and guide their verification.[9,10] Diagnostic performance depends therefore on the amount and richness of the illness scripts available in clinicians' memory. In the two aforementioned experiments, students who engaged in deliberate reflection were requested to match findings of the case with each alternative diagnosis.[7,8] This process probably fostered activation of relevant knowledge, establishment of connections and reorganisation of pre-existing knowledge, enriching the mental representations of the diseases and making it easier for these students to recognise new examples of the diseases in the test.

In these previous studies, the students reflected upon the cases without any guidance regarding which alternative diagnoses should be considered or regarding the features associated with them. A subsequent experiment investigated whether deliberate reflection could be made even more effective by providing additional instructional guidance.[11] It compared the effects of free reflection, cued reflection and modelled reflection. Whereas under free reflection students reflected on the cases without any guidance, under the cued-reflection condition, students were informed which diagnoses they should consider. Under the modelled-reflection condition, the highest degree of guidance, students were given a fully worked-out example of an expert's reflection upon the case and requested to study it. By 'focusing' reflection during practice, eliminating the possibility that students reflected upon unrelated diseases or failed to reflect upon plausible alternative diagnoses that would be tested subsequently, cued reflection was expected to foster performance relative to free reflection. Modelled reflection was expected to outperform cued reflection based on research on example-based learning.[12,13] This research has shown that studying examples of problems solved by a model leads to better learning outcomes (is more effective), often in less time and with less effort (i.e. is more efficient) than actually solving the same problems. This finding, demonstrated in tasks in several domains,[14–16] is attributed to the reduction of ineffective cognitive load.[17,18] Because students do not need to invest limited mental resources to search for the steps to solve the problem, these resources can be allocated to processing the information that is relevant for learning (for instance, features associated with the illness scripts).

However, the findings of the aforementioned experiment[11] were only partially in line with these expectations. Although the modelled reflection indeed performed better than the free-reflection group in the test, cued reflection and modelled reflection had similar performance, although the latter required less effort (i.e. was more efficient).[19] This similar performance occurred despite a potential advantage of the modelled-reflection condition, which is the exposure to additional knowledge provided by the worked-out example.

These unexpected findings may be explained by the expertise reversal effect, the phenomenon showing that studying examples, although advantageous for novices, becomes less effective as expertise increases.[20,21] This is because schemas storing domain knowledge in experienced learners' memory can be mobilised during problem solving and help organise incoming information. The students in the previous experiment[11] were in the final years of the curriculum and perhaps already had sufficient prior knowledge to provide scaffolding for the reflection process. If this were true, the advantage of studying examples of reflection over actually practising reflection, even if supported by cues, would show up when more novice students were involved.

To test this idea, the present study replicated (in an extended version) the previous experiment with novice students. Third-year students practised the diagnosis of clinical cases by engaging in free reflection, cued reflection or modelled reflection. Two weeks later, a diagnostic test was administered that consisted of new examples of the same diseases seen in the learning phase (hereafter 'studied diseases') and new diseases that were not seen during the learning phase but were plausible diagnoses for the studied cases (hereafter 'adjacent diseases'). We hypothesised that modelled reflection would lead to better performance in the test and require less investment of mental effort relative to both cued and free reflection; cued reflection was expected to be more beneficial than free reflection, although demanding similar or less mental effort. This would only apply, however, to the studied diseases. Because only students from the free-reflection condition could have reflected upon the adjacent diseases in the learning phase (they had to generate the alternative diagnoses by themselves), they would possibly perform better on the adjacent diseases in the test.

METHODS

Setting and participants

The participants were 80 third-year medical students (mean age = 21.23 years, standard deviation = 2.07; 43 male) from the Federal University of Minas Gerais, Belo Horizonte, Brazil. The school has a 6-year curriculum, with clerkships in the last 18 months. We chose third-year students because they had attended the courses addressing the study diseases in the previous term and, without any clinical experience, could be considered novices. All third-year students (160) were invited by three co-authors (TF, SMES and RMDdF) to voluntarily participate in the study. Those who accepted and completed both phases of the study were included as participants. Written consent was obtained from all participants, who did not receive any compensation.

The Research Ethics Committee of the Faculty of Medicine, Federal University of Minas Gerais provided approval for the study (CAAE # 48555915.4.0000.5149).

Material and procedure

The study used 19 written clinical cases, each one consisting of a brief description of a patient's medical history, present complaints, findings from physical examinations and results of diagnostic tests. Table 1 presents an overview of the cases' diagnoses. Filler cases were included in both phases to avoid participants easily recognising a pattern in the diagnoses.

The cases were selected from a collection of cases used in previous studies with students from the same city.[7,8,11] Based on diagnostic accuracy scores obtained in these studies, it was possible to allocate more difficult cases for the test, making the diagnostic task in the test more demanding. The cases were presented to participants in a booklet. Two versions of the booklets were prepared, alternating the order in which the cases were presented; these versions were randomly distributed to the students in each session.

Learning phase

The learning phase consisted of diagnosing nine clinical cases (Table 1) by following different instructions depending on the experimental condition to which the students were randomly assigned. For all conditions, the first page of each case presented the case description, and the student was required to read the case and write down the most likely diagnosis. On the second page, the case was presented again, and the instructions differed for each condition. Under the free-reflection condition, following the deliberate reflection procedure,[5,7,8] the students were requested to fill in a table that required them to: (i) list the findings in the case that

Table 1 Overview of the diagnoses of the cases used in the study

Learning phase Criterion cases	Diagnostic test Criterion cases
Acute viral hepatitis	Acute viral hepatitis*
Choledocholithiasis	Choledocholithiasis*
Alcoholic cirrhosis	Acute myocardial infarction*
Acute myocardial infarction	Acute viral pericarditis*
Acute viral pericarditis	Pancreas carcinoma†
Aortic dissection	Haemolytic anaemia†
	Chest wall pain†
	Gastro-oesophageal reflux†
Filler cases	**Filler cases**
Community-acquired pneumonia	Meningoencephalitis
Nephrotic syndrome	Infectious mononucleosis
Visceral leishmaniasis	

* Studied diseases.
† Adjacent diseases.

support the initial diagnosis; (ii) list the findings that speak against this diagnosis; (iii) list the findings that were expected to be present if this diagnosis was correct but are absent; (iv) list alternative diagnoses, if the initial diagnosis was incorrect, and (v) follow the same procedure (steps 1–3) for each alternative diagnosis. Finally, they were asked to rank the diagnoses in order of likelihood, thereby deciding on the most likely diagnosis for the case. Under the cued-reflection condition, the instructions on the second page of the case were the same, except that the table used for reflecting on the case already included the alternative diagnoses. The students were requested to complete the cells of the table and to rank the diagnoses in order of likelihood. Finally, under the modelled-reflection condition, the second page of the case presented the fully completed table, including the likelihood of the diagnoses, and the students were instructed to study the table.

For each case, the students were asked to rate how much mental effort they had to make to give the diagnosis (on the first page of the case) and to carry out the different forms of reflection (on the second page of the case). They did so by using a well-validated 9-point scale[22] ranging from 1 (very, very low mental effort) to 9 (very, very high mental effort), which was included at the bottom of each page.

Time allocated for each page of the case was the same under all conditions. In line with previous studies,[7,8,11] 1.5 minutes was allocated for the first page (provide the diagnosis) and 5.5 minutes for the second page (reflection on the case). Throughout the session, a researcher (TF-S, SMES, RMDdF) controlled the time, indicating when the students should turn each page.

Diagnostic test

Two weeks after the learning phase, a diagnostic test required all students to diagnose 10 new clinical cases (eight criterion cases; two fillers). The criterion cases consisted of four new exemplars of diseases studied in the learning phase and four cases of adjacent diseases. (See Table 1).

The students were requested to read about the case and provide the most likely diagnosis. As in the learning phase, they indicated the mental effort required to diagnose each case by using the 9-point scale. Time to work on each case was not controlled, but participants were informed that a maximum time of 50 minutes was allocated for the test. After having diagnosed all cases, students provided demographic information and answered (Yes or No) questions on whether they felt motivated to study the diseases seen in the learning phase and whether they actually studied them in between the learning phase and test.

One month after the test, students were debriefed about the study and received feedback on the diagnosis of the clinical cases in a session offered by expert internists.

Data analysis

The diagnoses provided by the participants for each case were evaluated using a 3-point scale of incorrect, partially correct or correct and scored, respectively, as 0, 0.5 or 1. A response was considered correct when the core diagnosis of the case was mentioned (e.g. myocardial infarction for the case of acute myocardial infarction). When only a component of the correct diagnosis was written, the response was evaluated as partially correct (e.g. myocardial ischaemia for the same case). Responses that did not fall into either of these categories were evaluated as incorrect.

This 3-point scale was employed in previous studies by board-certified internists who evaluated the participants' responses to the same clinical cases.[7,8,11] The scoring grids generated in the previous studies were used to score the participants' responses in the present study. First, the responses were transcribed to tables containing all the responses given to each case without displaying the experimental condition under which they had been provided. Subsequently, three co-authors (SM, TF-S and SMES) independently matched each response given in the present study to responses provided for the same case in previous studies. The score attributed in previous studies to an equal (or paraphrased) response was then assigned to the response provided in the present study. The raters agreed on the scores attributed to 98% of the responses and solved discrepancies through discussion.

For each participant, we summed the scores obtained in the learning phase for the four diseases that appeared in all phases (studied diseases). A mean diagnostic accuracy score (range 0–1) was computed for each participant and subsequently for each experimental condition. In a similar way, we computed the mean diagnostic accuracy score obtained in the test on the studied diseases and on the adjacent diseases.

For each participant, mental effort ratings on the first task (providing the diagnosis) and the second task (reflecting on the case) of the learning phase, as well as on the test items, were averaged separately, and mean effort ratings for diagnosing, reflecting and completing the test were computed for each experimental condition.

Firstly, we checked whether there were prior differences between the groups that could potentially influence performance in the test. A chi-squared test was performed to check similarity in gender distribution, and separate one-way Analyses of variance (ANOVAs) with experimental condition (free reflection, cued reflection or modelled reflection) as a between-subjects factor were performed on students' age, academic achievement rating (a score based on the student's grades in the previous term, ranging from 0 to 5, obtained from the regular school information system and matched to the participants' data) and mean diagnostic accuracy score obtained in the learning phase. Chi-squared tests compared students' motivation to study the diseases and their actual engagement in study after the learning session.

Two separate one-way ANOVAs with experimental condition as a between-subjects factor were performed on the diagnostic accuracy scores obtained in the test on the studied diseases and on the adjacent diseases. Post hoc tests (with Šídák correction) were used to further explore a significant main effect of experimental condition.

A mixed ANOVA with experimental condition as between-subjects factor and performance moment (learning phase, first task; learning phase, second task; test phase) as a within-subjects factor was performed on the mean ratings of mental effort reported at each moment. Significant effects were further analysed by performing one-way ANOVAs and post hoc tests (with Šídák correction).

If mental effort 'accumulates' throughout the task, it may become so high that it disturbs performance, especially for those making more effort. We, therefore, examined the three cases solved first and the three cases solved last in the learning session separately (note that case difficulty is not an issue because the different versions of the booklets ensured counterbalancing). We computed mean diagnostic performance score and mental effort ratings on these two 'types' of cases and performed separate ANOVAs with experimental condition as a between-subjects factor and performance moment (first-solved cases and last-solved cases) as a within-subjects factor. Significant effects were further explored by performing one-way ANOVAs and post hoc tests (with Šídák correction).

The significance level was set at $p < 0.05$ in all analyses. SPSS Version 25 (IBM Corp., Armonk, NY, USA) for Mac was used for the statistical analysis.

RESULTS

Participants' characteristics

Table 2 presents the participants' characteristics. The groups did not differ in gender ($x^2(2) = 0.84$; $p > 0.6$) or age ($F_{(2,77)} = 1.34$; $p > 0.2$). Academic achievement ratings were also statistically similar ($F_{(2,77)} = 2.82$; $p > 0.05$). The percentage of students who reported they were motivated to study the diseases ($x^2(2) = 0.56$; $p > 0.7$) or who actually studied them after the learning session

($x^2(2) = 1.89$; $p > 0.3$) did not differ between the groups.

Diagnostic accuracy

The diagnostic accuracy scores obtained in the two phases are presented in Table 3. The groups performed similarly in the learning phase ($F_{(1,77)} = 2{,}36$; $p > 0.05$; $\eta_p^2 = 0.05$). In the test, there was a significant effect of experimental condition on diagnostic accuracy for the studied diseases ($F_{(2,77)} = 4.31$; $p < 0.05$; $\eta_p^2 = 0.10$). The cued-reflection group performed significantly better than the free-reflection group ($p < 0.05$). The difference between the modelled-reflection group and the free-reflection group was not statistically significant ($p > 0.05$), nor did a significant difference emerge between the cued-reflection group and the modelled-reflection group ($p > 0.05$). Concerning the adjacent diseases, there was no significant effect of experimental condition on the diagnostic accuracy scores, with the groups performing similarly ($F_{(2,77)} = 0.55$; $p > 0.05$; $\eta_p^2 = 0.01$).

Table 2 Participants' background characteristics and frequency of students who reported motivation to study and who did study the diseases seen in the learning phase before the test

	Free reflection *n* = 26 (SD)	**Cued reflection *n* = 26 (SD)**	**Modelled reflection *n* = 28 (SD)**
Gender (female; male)	13; 13	13; 13	11; 17
Age, years	21.23 (1.58)	22.00 (2.10)	22.07 (2.40)
Academic achievement rating (0–5)	4.10 (0.45)	3.78 (0.48)	3.90 (0.49)
Students who reported motivation to study the diseases	16	19	20
Students who engaged in study of the diseases	3	6	3

SD = standard deviation.

Table 3 Mean diagnostic accuracy scores (range 0–1) obtained in the learning phase and in the test as a function of experimental condition

	Free reflection *n* = 26 (SD)	**Cued reflection *n* = 26 (SD)**	**Modelled reflection *n* = 28 (SD)**
Performance in the learning phase	0.56 (0.20)	0.65 (0.21)	0.69 (0.25)
Performance in the test phase on the studied diseases	0.41 (0.20)	0.58 (0.23)	0.54 (0.22)
Performance in the test phase on the adjacent diseases	0.18 (0.17)	0.14 (0.18)	0.19 (0.21)

SD = standard deviation.

Mental effort

Table 4 presents the mean ratings of mental effort throughout the study. There was a significant main effect of performance moment ($F_{(2,150)} = 26.13$; $p < 0.001$; $\eta_p^2 = 0.26$) and a significant main effect of experimental condition ($F_{(2,75)} = 11.82$; $p < 0.001$; $\eta_p^2 = 0.24$). The interaction effect was also significant ($F_{(4,150)} = 16.06$; $p < 0.001$; $\eta_p^2 = 0.30$), with a different pattern emerging across time. In the learning phase, effort invested in providing the initial diagnosis differed between the groups ($F_{(2,76)} = 7.05$; $p < 0.01$; $\eta_p^2 = 0.16$), with the free-reflection group reporting more effort relative to both the cued-reflection group ($p < 0.01$) and the modelled-reflection ($p < 0.01$), with similar ratings being reported by the cued-reflection group and the modelled-reflection group ($p > 0.05$). A significant difference between groups also emerged in the reflection task ($F_{(2,76)} = 25.53$; $p < 0.001$; $\eta_p^2 = 0.40$). Although the mental effort ratings of the free-reflection group and the cued-reflection group did not differ significantly ($p > 0.05$), they were significantly higher than the mental effort ratings of the modelled-reflection group (both p-values < 0.001). On the test, however, the groups did not differ significantly in mental effort ratings ($F_{(2,77)} = 2.70$; $p > 0.05$; $\eta_p^2 = 0.06$).

Figure 1 presents the results of the post hoc exploratory analysis of developments in performance

Table 4 Mean ratings of mental effort (range 0–9) reported for the studied diseases in the two tasks of the learning phase and in the test and for the adjacent diseases in the test as a function of experimental condition

	Free reflection $n = 26$ (SD)	Cued reflection $n = 26$ (SD)	Modelled reflection $n = 28$ (SD)
Mental effort in the learning phase, first task (providing a diagnosis)	6.01 (1.06)	4.89 (1.33)	5.04 (1.17)
Mental effort in the learning phase, second task (reflecting on the case)	6.46 (1.28)	6.31 (1.14)	4.11 (1.58)
Mental effort in the test phase	6.73 (0.95)	6.11 (1.25)	6.11 (1.15)

SD = standard deviation.

and effort in the first and last third of the learning phase. Concerning diagnostic performance, there was a significant interaction effect ($F_{(2,77)} = 3.26$; $p < 0.05$; $\eta_p^2 = 0.08$), with the groups obtaining similar diagnostic accuracy scores on the initial diagnosis in the first three cases (all p-values > 0.05) but the free-reflection group performing more poorly than both the cued-reflection group and the modelled-reflection group (both p-values < 0.001) in the last cases solved in the session. An interaction effect was also found in the mental effort ratings ($F_{(2,75)} = 3.50$; $p < 0.05$; $\eta_p^2 = 0.09$). The groups reported similar effort required in performing the first task (providing the diagnosis) for the first three cases (all p-values > 0.05) but differed on the last three cases, with the free-reflection group requiring more effort to provide a diagnosis than both the cued-reflection group ($p < 0.01$) and the modelled-reflection group ($p < 0.05$).

DISCUSSION

The present study investigated whether it is beneficial to provide additional guidance in the form of worked examples when students learn clinical diagnosis through deliberate reflection. The findings only partially support our initial hypotheses. As expected, providing cues for reflection led to better performance in the test relative to free reflection when requiring a similar mental effort to reflect upon the cases in the learning phase. However, contrary to our prediction, studying examples of reflection was not more beneficial than actually engaging in (cued) reflection on the cases. Overall, the test scores of the modelled-reflection group did not differ significantly from those of either the cued-reflection group or the free-reflection group. It could be argued that modelled reflection was somewhat more efficient, as it demanded less effort in the learning phase, whereas test scores did not differ significantly from the other two conditions. By contrast, only the cued-reflection condition outperformed the free-reflection condition. That said, the exploratory case sequence analysis presented in Fig. 1 suggests that a benefit of worked examples relative to free reflection does emerge as the demands of testing increase (i.e. the longer one is engaged in the cognitively demanding task of diagnosing new cases). Any such differences in performance were limited to new examples of diseases studied in the learning phase; on the adjacent diseases there were no differences amongst groups and accuracy scores were very low.

Concerning the studied diseases, the findings replicate with novice students the results previously obtained with advanced students.[11] Simply cueing reflection so that it would focus on the 'relevant' diagnoses–diagnoses that would be addressed in the test–was sufficient to make students perform better than those who chose by themselves which diseases they would consider. Deliberate reflection involves matching findings from the case with each alternative diagnosis, which might help establish relationships between findings and restructure pre-existing knowledge of the disease under consideration. Cued reflection ensured the opportunity to reflect several times upon the same disease (with different clinical presentations), which would not necessarily happen when students generated alternative diagnoses by themselves. Not surprisingly, the minimal guidance provided by cued reflection improved performance relative to free reflection.

What is quite surprising is that modelled reflection did not improve performance relative to cued reflection, replicating with novices what was observed with advanced students.[11] These findings are unexpected not only in light of the existing evidence that studying examples is more beneficial

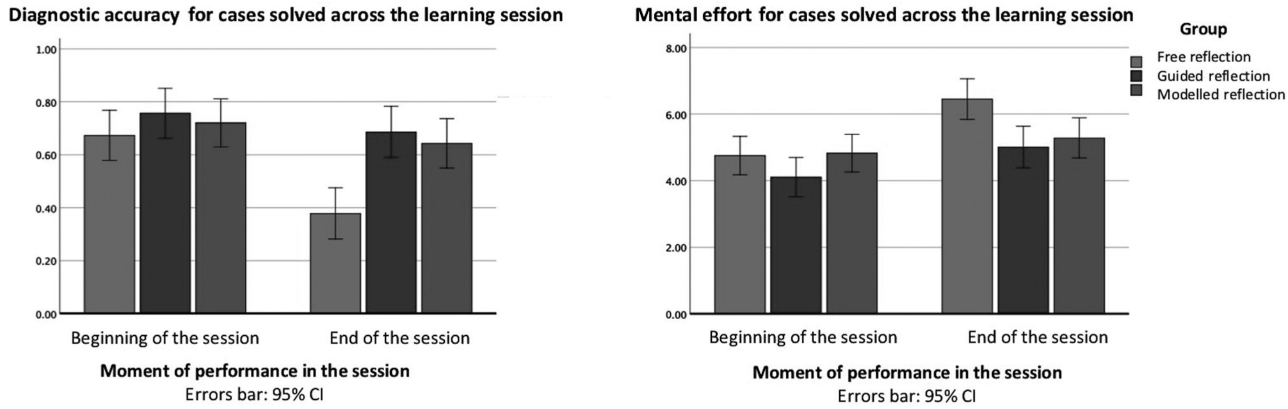

Figure 1 Diagnostic accuracy and mental effort in the first task of the cases by the start and by the end of the learning session. CI = confidence interval.

for novices than solving the equivalent problems,[13] but also because of the additional information provided by modelled reflection. When studying the experts' reflection models, students probably encountered information that was new to them. Conversely, the cued-reflection group and the free-reflection group were not exposed to new knowledge about the diseases.

What could explain why students did not benefit from modelled reflection as much as expected? It is known that studying examples becomes less effective (or even has a negative effect relative to problem solving) as learning progresses and learners can use schemas stored in long-term memory to process incoming information more efficiently.[20,21] Our participants had very limited expertise. Nevertheless, even their little prior knowledge was seemingly sufficient to allow the students who were cued during reflection to benefit from it to such an extent that they performed similarly to the modelled-reflection students despite the disadvantage of not having been exposed to new knowledge. The higher mental effort experienced by the cued-reflection students in the learning phase does not seem, therefore, to have been caused by ineffective cognitive processes (i.e. extraneous load).[23] It may be that the nature of medical knowledge, clustered around diseases, makes it possible that rudimentary schemas develop early in training. These nascent schemas would then allow even novice medical students to benefit from problem solving, reducing the advantage of studying examples, which is different to what has been found in research with ill-structured problems in other domains such as law.[24] This is, however, only a conjecture that demands investigation.

Regarding the adjacent diseases, all groups performed similarly–and very poorly–on the test. Only students who engaged in free reflection during the learning phase could have reflected on the adjacent diseases, becoming potentially more able to recognise the diseases in the test. However, a post hoc count showed that the adjacent diseases were rarely considered as possible diagnoses during reflection (indeed two were never mentioned). This suggests very little familiarity with these diseases, an idea corroborated by the extremely low diagnostic scores in all conditions.

The post hoc analysis exploring performance across the learning session revealed intriguing findings. The effort required from free-reflection students to perform the second task (reflection task) apparently accumulates over time and transfers to the first task (provide a diagnosis). Even if the first (diagnosis) task was the same for all conditions, the free-reflection group found it required more effort in the last part of the learning phase. This was not true, however, for the first three cases in the learning phase, when both diagnostic accuracy and mental effort were similar for all conditions. The higher cognitive load only emerged for cases diagnosed by the end of the session, at which point the performance of free-reflection students on the initial diagnosis task was also lower. This higher effort combined with lower performance may have been caused by exhaustion or by (short-term) confusion as a result of engaging in free reflection. This mirrors prior findings[7] that engaging in free reflection led to lower performance on a test that immediately followed the learning phase, yet better performance on a delayed test.

This study has potential limitations. First, the sample size was small so would not allow for detecting very small effects. We had a 2-week delayed test but have not checked whether the effect of the approaches would last further than this time lag. Finally, because of the unexpectedly high level of complexity of the adjacent diseases for the participants, we could not test whether leaving students free to reflect by themselves would have benefits in terms of transfer to new cases. These are issues to be addressed in future research.

CONCLUSION

The findings of this study may help clinical teachers in their choices of approaches to be used during students' practice with clinical cases. Evidence on instructional approaches to be adopted in such practice is much needed.[2,4] Deliberate reflection has been shown to be more effective than conventional approaches such as making a differential diagnosis.[7,8] The present study shows that providing even a little additional guidance, such as cues on which diseases to consider during reflection, has an added value, making deliberate reflection even more effective for novice students. Cueing deliberate reflection does not require additional effort from teachers. Studying examples of reflection is an option that requires less effort from students but preparing the worked examples demands much more investment from teachers.

Contributors: all authors (SM, TF-S, SMES, RMDdF, HGS and TvG) agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. Study concept and design: SM and TvG. Acquisition of data: TF-S, SMES, RMDdF. Statistical analysis: SM. Analysis and interpretation of data: SM, TF-S, SMES, RMDdF, HGS and TvG. Drafting of the manuscript: SM. Critical revision of the manuscript for important intellectual content: SM, TF-S, SMES, RMDdF, HGS and TvG. Administrative, technical or material support: TF-S, SMES, RMDdF and SM. Study supervision: SM and TvG.
Acknowledgments: the authors are grateful to the students who gave their time to participate in the study.
Funding: the study was supported by a grant provided by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES)/Edital Pró-Ensino da Saúde 1606/2011 (CAAE # 03896812.4.0000.5149), Ministry of Education, Brazil.
Conflicts of interest: none.
Ethical approval: ethical approval for the study was provided by the Research Ethics Committee of the Faculty of Medicine, Federal University of Minas Gerais (CAAE # 48555915.4.0000.5149).

REFERENCES

1 Norman G. Research in clinical reasoning: past history and current trends. *Med Educ* 2005;**39** (4):418–27.
2 Kassirer JP. Teaching clinical reasoning: case-based and coached. *Acad Med* 2010;**85** (7):1118–24.
3 Eva KW. What every teacher needs to know about clinical reasoning. *Med Educ* 2005;**39** (1):98–106.
4 Schmidt HG, Mamede S. How to improve the teaching of clinical reasoning: a narrative review and a proposal. *Med Educ* 2015;**49** (10):961–73.
5 Mamede S, Schmidt HG, Penaforte JC. Effects of reflective practice on the accuracy of medical diagnoses. *Med Educ* 2008;**42** (5):468–75.
6 Mamede S, van Gog T, van den Berge K, Rikers RM, van Saase JL, van Guldener C, Schmidt HG. Effect of availability bias and reflective reasoning on diagnostic accuracy among internal medicine residents. *JAMA* 2010;**304** (11):1198–203.
7 Mamede S, van Gog T, Moura AS, de Faria RM, Peixoto JM, Rikers RM, Schmidt HG. Reflection as a strategy to foster medical students' acquisition of diagnostic competence. *Med Educ* 2012;**46** (5):464–72.
8 Mamede S, van Gog T, Sampaio AM, de Faria RM, 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** (1):121–7.
9 Schmidt HG, Norman GR, Boshuizen HPA. A cognitive perspective on medical expertise - theory and implications. *Acad Med* 1990;**65** (10):611–21.
10 Schmidt HG, Rikers RM. How expertise develops in medicine: knowledge encapsulation and illness script formation. *Med Educ* 2007;**41** (12):1133–9.
11 Ibiapina C, Mamede S, Moura A, Eloi-Santos S, van Gog T. Effects of free, cued and modelled reflection on medical students' diagnostic competence. *Med Educ* 2014;**48** (8):796–805.
12 Van Gog T, Rummel N. Example-based learning: integrating cognitive and social-cognitive research perspectives. *Educ Psychol Rev* 2010;**22** (2):155–74.
13 Van Merrienboer JJ, Sweller J. Cognitive load theory in health professional education: design principles and strategies. *Med Educ* 2010;**44** (1):85–93.
14 Rummel N, Spada H. Learning to collaborate: an instructional approach to promoting collaborative problem-solving in computer-mediated settings. *J Learn Sci* 2005;**14**:201–41.
15 Couzijn M. Learning to write by observation of writing and reading processes: effects on learning and transfer. *Learn Instr* 1999;**9** (2):109–42.
16 Hilbert TS, Renkl A. Learning how to use a computer-based concept mapping tool: self-

explaining examples helps. *Comput Human Behav* 2009;**25**:267–74.
17 Sweller J, van Merrienboer JJG, Paas FGWC. Cognitive architecture and instructional design. *Educ Psychol Rev* 1998;**10** (3):251–96.
18 Van Merrienboer JJG, Sweller J. Cognitive load theory and complex learning: recent developments and future directions. *Educ Psychol Rev* 2005;**17** (2):147–77.
19 Van Gog T, Paas F. Instructional efficiency: revisiting the original construct in educational research. *Educ Psychol* 2008;**43** (1):16–26.
20 Kalyuga S, Chandler P, Tuovinen J, Sweller J. When problem solving is superior to studying worked examples. *J Educ Psychol* 2001;**93** (3):579–88.
21 Kalyuga S, Ayres P, Chandler P, Sweller J. The expertise reversal effect. *Educ Psychol* 2003;**38** (1):23–31.
22 Paas F. Training strategies for attaining transfer problem-solving skills in statistics: a cognitive load approach. *J Educ Psychol* 1992;**84**:429–34.
23 Sweller J, Ayres P, Kalyuga S. *Cognitive Load Theory. Explor Learn Sci.* 2011;New York, NY: Springer-Verlag New York, Inc. 3–261.
24 Nievelstein F, van Gog T, van Dijck G, Boshuizen HPA. The worked example and expertise reversal effect in less structured tasks: learning to reason about legal cases. *Contemp Educ Psychol* 2013;**38** (2):118–25.

Received 9 September 2018; editorial comments to authors 30 November 2018; accepted for publication 11 January 2019