# TEACHING DIAGNOSTIC REASONING IN MEDICAL TRAINING

## INVESTIGATING INTERVENTIONS TO TEACH DELIBERATE REFLECTION AND IMPROVE DIAGNOSTIC CALIBRATION

JOSEPHA KUHN

**Colophon**

Lay-out: Robert Koopman
Print: De Boekdrukker, Amsterdam



This research was funded by ZonMW, The Netherlands (839130007).

# Teaching Diagnostic Reasoning in Medical Training - investigating interventions to teach deliberate reflection and improve diagnostic calibration

Leren van diagnostisch redeneren in de medische opleiding -
Onderzoek naar interventies voor het aanleren van de weloverwogen reflectieprocedure
en het verbeteren van diagnostische kalibratie

## Thesis

to obtain the degree of Doctor from the
Erasmus University Rotterdam
by command of the
rector magnificus

Prof. Dr. A.L. Bredenoord

and in accordance with the decision of the Doctorate Board.
The public defence shall be held on

Wednesday 17 May 2023 at 13:00 hrs
by

**Josepha Katharina Gisela Kuhn**
born in Magdeburg, Germany

**Erasmus University Rotterdam**

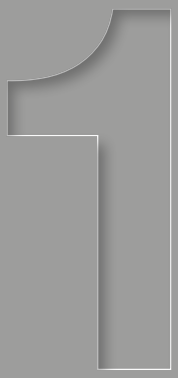
**Doctoral Committee**

**Promotors:**
Prof. Dr. T.A.J.M. van Gog
Prof. Dr. P.J.E. Bindels

**Other members:**
Prof. Dr. W.W. van den Broek
Prof. Dr. F.G.W.C. Paas
Prof. Dr. A.B.H. de Bruin

**Copromotors:**
Dr. S. Mamede Studart Soares
Dr. P.J. van den Berg

**TABLE OF CONTENTS**



# CHAPTER 1

## General Introduction

The purpose of this dissertation was to study whether and how reflective diagnostic reasoning could be learned by general practitioners in training and how we could stimulate them to apply it. Three studies focussed on the question of whether a procedure to analyse clinical cases, called *deliberate reflection*, could be taught to general-practice residents and medical students. A fourth study investigated whether we could help residents to better estimate how well they have diagnosed a case, which is important for recognising when a case may need more attention and applying reflective reasoning may be beneficial for solving the case. The written cases used in these studies always described patient encounters as they could happen in a Dutch general practice. In the Netherlands like in some other European countries, the general practitioner (GP) is always the first one to see a patient. The general practitioner then, has to assess how serious the case is, while having only limited access to additional diagnostic tests when compared to a hospital setting. While primary care has to deal with diagnostic uncertainty (Bhise et al., 2018), general practitioners still have to make important decisions, for example whether it is appropriate to wait and see how the symptoms will develop, whether treatment should be started, or whether the patient needs to be referred to a hospital specialist. This makes good diagnostic reasoning skills crucial in general practice. A study in Dutch general practices found that 2,1% of the patient records showed patient safety incidents related to the diagnosis (Gaal et al., 2011). Studies from other disciplines indicate that diagnostic error are often related to cognitive errors (Graber et al., 2005), which means that improving diagnostic reasoning skills may improve patient safety. In this Introductory chapter, I will first present the theoretical background for the studies we conducted. Then, I will give an overview of this dissertation with a description of each study and the research questions.

## Diagnostic Reasoning

Diagnostic reasoning is the cognitive process that leads to a diagnosis for a clinical case. In that process, a physician categorises the signs and symptoms from the case and compares them with previous knowledge and experience with different medical conditions, which can lead to a diagnosis (Higgs & Jones, 2000; Patel et al., 2013). This diagnosis should then be further tested (Kuhn, 2002). The underlying cognitive processes that lead to the diagnosis do differ between physicians, depending on their level of experience and the difficulty of the case (Elstein & Schwartz, 2002; Schmidt & Boshuizen, 1993; Schmidt et al., 1990; Schmidt & Rikers, 2007). When a novice, for example a medical student, diagnoses a case, each symptom is consciously analysed. As a physician gains more experience and sees more and more cases, the knowledge gets structured differently. The knowledge of all the different cases of the same disease that the physician has encountered gets encapsulated into so called *illness scripts*, which are mental models of a disease that combine all the clinically relevant information about the disease into one concept. If good illness scripts have been

formed, a physician can diagnose a case by pattern recognition, instead of analysing each single feature. The case at hand can then quickly be recognised as being similar to cases encountered previously.

This means that the development of medical expertise involves not only gaining knowledge, but also restructuring of that knowledge, which enables experts to come to good diagnoses quickly. These diagnoses can then be further tested. Whenever the diagnosis turns out to be wrong, or physicians cannot clearly recognise a disease pattern, they can switch back to a more analytical way of diagnosing (Mamede et al., 2007). This means that expert physicians can switch between different strategies to diagnose a case and none of the two ways of reasoning is preferred to the other as they can both lead to correct diagnoses (Schmidt et al 1990).

This theory, that experts can switch between pattern recognition and analytical reasoning, is in line with dual-process theories of thinking. Several versions of this theory have been described (Evans, 2008) which all have in common that they describe two different modes of processing. *System 1* processing describes a non-analytical, automatic, unconscious, effortless way of thinking. In this mode, people make use of pattern recognition and heuristics, which can lead to fast decision making and therefore is very efficient. *System2* processing describes an analytical, reflective, and conscious way of thinking. Both ways of processing can lead to correct conclusions and decisions, and with both types processing errors can occur (Norman et al., 2017).

While diagnostic errors can have many causes, like system related errors or knowledge deficits, reasoning errors are thought to be an important contributor (Braun et al., 2017; Graber et al., 2005; Kuhn, 2002; Zwaan et al., 2012). This means that in cases where physicians had the necessary knowledge to come to the correct diagnosis and the necessary information from the case was available, reasoning errors lead to a wrong or delayed diagnosis.

One source of reasoning errors lies in cognitive biases (Kahneman, 2011). In the medical field, many potential cognitive biases have been identified that can unconsciously influence a physician and lead to a diagnostic error (Croskerry, 2003). For example, when physicians diagnose a case that resembles a case they have recently diagnosed (Mamede, Van Gog, et al., 2010), or resembles a diagnosis they have recently read about (Schmidt et al., 2014), they are more likely to give the same diagnosis based on System 1 processing, even if it is wrong for the case at hand. This is known as availability bias. Diagnostic errors can also be related to confirmation bias, where physicians engage in System 2 processing to some extent, but only actively look for information that supports their diagnosis and overlook or give

less weight to information that would contradict it (Mendel et al., 2011; Zwaan et al., 2013). While System 1 processing is highly efficient, and while errors can occur with both types of processing, it is important in many situations to engage in System 2 processing and deliberately analyse the case to be able to prevent or correct errors (Mamede, Van Gog, et al., 2010). Also, when cases are complex, consciously thinking about it (as opposed to relying on pattern recognition) can help with solving the case, at least for experts who can then apply their medical knowledge (Mamede, Schmidt, et al., 2010).

Because improving diagnostic reasoning may help physicians to avoid diagnostic errors and may improve patient safety, many attempts have been made to develop interventions for improving diagnostic reasoning, with varying effectiveness. Among the few that have shown a positive effect on diagnostic accuracy, are interventions based on reflection (Griffith et al., 2021; Lambe et al., 2016; Prakash et al., 2019).

## Deliberate Reflection

Different interventions to reduce diagnostic error by fostering reflection have been tested. For example, some studies have given general instructions to be more reflective when diagnosing a case (Sibbald & de Bruin, 2012) or to simply have a second look (Monteiro et al., 2015); others have used educational sessions to explain cognitive biases and the importance of reflection (Sherbino et al., 2014); or have implemented different types of checklists to improve diagnosis (e.g., a differential diagnosis checklist after having given the initial diagnoses intuitively, or more general metacognitive checklists; Chew et al., 2016; Ely et al., 2011; Shimizu et al., 2013). Studies on these interventions differ in how their effectiveness has been evaluated and whether diagnostic accuracy has been assessed. However, recent reviews have shown that interventions that provide detailed step-by-step instructions to reflect on the initial diagnosis have been most consistently successful to prevent bias and improve diagnostic accuracy (Griffith et al., 2021; Lambe et al., 2016; Prakash et al., 2019). Most of these studies employed *deliberate reflection* (Mamede, Schmidt, & Penaforte, 2008), which is also the focus of this dissertation.

Deliberate reflection is a procedure to reflect on a clinical case, by following a set of questions that help to systematically analyse the features from the case in relation to different possible diagnoses. In studies on deliberate reflection, physicians get a written clinical case, read it, and are asked to write down their diagnosis for the case. After that, they are presented with a table as shown in Figure 1. They write down their diagnosis in the first free row and then answer the following questions by listing (present and absent) features from the case: (1) Which findings from the case speak for this diagnosis? (2) Which findings from the case speak against this diagnosis? (3) Which further findings would you expect if this diag-

nosis were true that are absent in the case? Then, they are asked to think of an alternative diagnosis, write it down in the empty row below and again answer the reflective questions for this diagnosis. If possible, repeat the procedure for a third potential diagnosis. Only after having analysed several diagnoses following these steps, physicians are asked to rank the diagnoses in order of likelihood and come to a final decision.



Figure 1 - Example of a table that can be filled in when diagnosing a case to guide a physician through the deliberate reflection procedure.

| Diagnosis | Which findings from the case speak **for** this diagnosis? | Which findings from the case speak **against** this diagnosis? | Which further findings would you expect if your diagnosis were correct, which are **absent** in the case? | Finally: Order of likelihood (1 = most likely) |
|---|---|---|---|---|
| | | | | |
| | | | | |
| | | | | |

The aim of these steps is to help physicians to consciously reflect on the case (to engage in System 2 thinking) which may help to correct errors that have been made with pattern recognition (System 1). It is also designed to help to counteract confirmation bias and get physicians out of a tunnel vision formed by their first impression of the case, as it asks physicians to actively look for information that is absent or speaks against their diagnosis, as well as to analyse several possible other diagnoses. Last but not least, this procedure may also counteract premature closure, where a physician accepts a diagnosis without fully verifying it (Croskerry, 2003), which occurs with physicians of all levels of expertise (Braun et al., 2017; Graber et al., 2005; Voytovich et al., 1985). Research has shown that deliberate reflection can indeed help to correct initial diagnostic errors when physicians were influenced by availability bias (Mamede, Van Gog, et al., 2010), when they were distracted by disruptive behaviour of the patient (Schmidt et al., 2017), when they were misled by some salient features of the case (Mamede et al., 2012), or when cases were complex and automatic pattern recognition may have failed them (Mamede, Schmidt, & Penaforte, 2008).

## Teaching Deliberate Reflection

In the above-mentioned studies that showed a benefit of deliberate reflection, physicians were always actively asked to apply the procedure and fill in the table. This is not a very practical or feasible procedure in clinical practice, however. For deliberate reflection to contribute to error reduction in practice, physicians would need to be able to rapidly and au-

tonomously apply it when diagnosing cases. Therefore, **the first aim of this dissertation** was to investigate whether we could teach physicians in training the deliberate reflection procedure and whether they would then apply it autonomously when diagnosing new cases later on.

Researchers in medical education have been debating whether diagnostic reasoning can be learned as a general skill that can then be applied to different kinds of medical problems, or whether it is content specific and does not transfer to solving different problems (Eva et al., 1998; Monteiro et al., 2020). Interventions that focus on changing physicians' reasoning process are often not effective to improve diagnostic accuracy (Norman et al., 2017; Schmidt & Mamede, 2015), or to protect against cognitive bias (Sherbino et al., 2014), which suggests that it cannot be learned as a general skill. Others have argued that transfer of a reasoning strategy to different problems is difficult but not impossible to achieve (Eva et al., 1998) and could help prevent cognitive bias (Croskerry et al., 2013).

This debate is not limited to medical education. Whether a cognitive strategy should be taught as a general skill or linked to specific content is also discussed with regard to teaching critical thinking more generally. Two meta-analyses have found that both approaches can show some effect, but a combined approach is the most effective (Abrami et al., 2015; Abrami et al., 2008). The best way to improve students' critical thinking skills is to teach the general critical thinking principles and then also teach how to apply these principles in a domain-specific content. The same may be true for teaching reflective reasoning in medicine. It might be most effective to teach the general procedure explicitly, but to also demonstrate how it can be applied to specific content (i.e., cases).

## Instructional Approaches for Teaching Deliberate Reflection

Educational research, and more specifically, instructional design research, has yielded insight into effective instructional approaches for teaching novices problem-solving and reasoning skills (Van Merriënboer & Sweller, 2010). In this dissertation, two approaches are investigated, also in combination, that have been shown to be effective learning various kinds of skills, although their effectiveness for teaching deliberate reflection has not yet been tested: Example-based learning (Van Gog et al., 2019) and learning-by-teaching (Duran, 2017).

**Example-based learning**. In example-based learning, novices learn how to solve a task by studying examples that demonstrate how the task could or should be solved. This has been shown to reduce extraneous (i.e., ineffective) working memory load when compared to performing the task themselves (*learning-by-doing*) and therefore more of the working memory capacity can be directed at activities that foster learning and transfer (Paas & Van

Gog, 2006). Consequently, example-based learning has been found to be a more effective (i.e., higher learning outcomes) and efficient (i.e., higher learning outcomes attained in less time and/or with less effort) method than learning-by-doing for acquiring problem-solving skills in various domains (Atkinson et al., 2000; Van Gog et al., 2019) including medical diagnosis (Stark et al., 2011). In addition, it has also been used to teach other types of skills, such as collaboration (Rummel & Spada, 2005) and self-regulation (Kitsantas et al., 2000).

Example-based learning has also been successfully used in combination with the deliberate reflection procedure, as a means to acquire diagnostic knowledge of certain diseases (Ibiapina et al., 2014). In that study, students diagnosed cases either by following the steps of deliberate reflection themselves (free reflection), by being given the diagnoses which should be analysed with deliberate reflection (cued reflection), or by studying full worked-out examples of how an expert used deliberate reflection to solve the cases (example-based learning). Participants also rated their mental effort investment during this learning session. In an immediate test and a delayed test one week later, the students diagnosed new cases of the same diseases they had seen in the learning phase. Cued reflection and example-based learning were equally effective for improving diagnostic accuracy in the test sessions and outperformed free reflection. Example based learning, however, required significantly less mental effort. While this study underlines the effectiveness and efficiency of example-based learning for acquiring diagnostic knowledge, it remained unclear (i.e., it was not directly tested) whether students had actually learned the deliberate reflection procedure from observing examples or engaging in cued reflection and could apply it on later cases.

**Learning-by-teaching**. In order to achieve transfer of the knowledge or skill to new problems, it can be helpful to engage with the learning material more actively (Brown & Kane, 1988; Eva et al., 1998; Van Gog & Rummel, 2010). An extension of example-based learning that can increase this active engagement and help learners to focus on the important parts of the learning material is *learning-by-teaching* (Lachner et al., 2021). Learning-by-teaching entails that learners are first asked to study learning content (e.g., examples), and then explain what they have learned to a (fictitious) peer. Teaching the material is a generative learning activity that can stimulate deeper processing than only studying the examples, which is beneficial for learning (for a review of generative learning activities: Fiorella & Mayer, 2016). It has been found that learning can improve even when students are merely expecting that they will be asked to teach the material after their study session, but actually teaching it improves learning even more, has a longer lasting effect, and helps with transfer to other problems (Fiorella & Mayer, 2013, 2014; Hoogerheide et al., 2014; Lachner et al., 2021). When students were video recorded while teaching, it was found that they experienced more arousal and perceived more cognitive load than when restudying or summarising the material (Hoogerheide et al., 2016; Hoogerheide et al., 2019). The feeling of a social pres-

ence (i.e., the fictitious audience) when being video recorded and considering this potential audience when explaining are also thought to contribute to the improvements in learning and transfer. These teaching methods may also be effective for physicians in training to learn deliberate reflection as a procedure for diagnostic reasoning.

## Knowing when to apply reflection: Improving diagnostic calibration

While asking physicians to engage in deliberate reflection can help with diagnostic accuracy, physicians do also switch to a more reflective reasoning approach naturally when cases are perceived as difficult (Mamede, Schmidt, Rikers, et al., 2008). It has been found, however, that physicians are not always good at estimating whether further reflection is needed (Monteiro et al., 2015) or how well they have diagnosed a case (Davis et al., 2006). It is important for physicians to recognise when their diagnosis is not yet correct and a case requires further attention. Therefore, **the second aim of this dissertation** was to investigate whether we could help physicians to better estimate their diagnostic performance after diagnosing a case.

Inadequate *diagnostic calibration*, which is a measure of the extent to which the diagnostic accuracy and the physician's confidence in the diagnosis are aligned, can lead to diagnostic error. When physicians are too confident in a wrong diagnosis, they may stop looking for alternative explanations (premature closure), which would hinder recognising a diagnostic error (Berner & Graber, 2008). Research has found that physicians do indeed tend to be overconfident in their diagnosis (Costa Filho et al., 2019; Friedman et al., 2005; Meyer et al., 2013). If physicians are, on the other hand, underconfident in a correct diagnosis, this may unnecessarily lengthen the diagnostic process including unnecessary medical testing. This means that physicians have to decide whether their current diagnosis is correct with the risk of undertesting if it were incorrect, or decide that the diagnosis is incorrect with the potential of overtesting an already correctly diagnosed case, which is a difficult balancing act (Meyer & Singh, 2019). Besides that, being able to correctly assess one's own diagnostic performance may help with recognising where one's learning needs are. Therefore, it may help physicians in training to make good use of their education and it may foster physicians' lifelong learning (Davis et al., 2006; Hacker & Bol, 2019).

Research from cognitive psychology has studied factors and interventions that could help to improve calibration. For example, some studies have found that when students got feedback on their performance, it helped them to make better self-assessments in the future. Feedback helped students to become better calibrated when recalling word definitions (Lipko et al., 2009; Nederhand et al., 2019) or when solving mathematical problems (Labuhn et al., 2010). In these studies, students who performed worse were usually also less

calibrated (they were more overconfident) and benefitted the most from the intervention. It has been suggested, that feedback may also help to improve diagnostic calibration in physicians (Meyer & Singh, 2019). A study by Nederhand et al. (2018) tested this and found that indeed, when medical students and experts got feedback on their previous performance, their diagnostic calibration for following cases improved. The cases used in that study were easy cases, however, and it has been found that physicians have more difficulty to judge their performance for difficult cases where they make more diagnostic errors (Meyer et al., 2013). Therefore, we tested whether feedback on previous diagnostic performance would help physicians to better estimate their diagnostic performance on future cases, when these cases were more difficult ones.



## Overview of this Dissertation

Chapters 2 – 4 describe three empirical studies (i.e., one per chapter) addressing the first aim of this study: To test whether deliberate reflection can be taught to physicians in training, so that they would apply it on future cases without instructions to reflect. In these studies, we tested different interventions to teach deliberate reflection, different ways to measure reflection, whether general-practice residents or medical students would adopt the procedure and whether they would apply it under specific circumstances. Chapter 5 describes an empirical study addressing the second aim of this study: To test whether we could help physicians in training to better estimate their performance after diagnosing a case. All studies were experimental studies conducted with written cases, describing patient encounters in a general-practice setting.

The first study, described in **Chapter 2**, compared two approaches to teaching deliberate reflection to general-practice residents: learning-by-doing and example-based learning. The study consisted of three sessions: a learning session and two test sessions. In the learning session, participants were randomly assigned to one of three conditions. They learned deliberate reflection via learning-by-doing or example-based learning, or they did not learn deliberate reflection (control condition). In the first test session a couple of hours later and the second test session a week later, all participants took the same test in which they diagnosed new cases and explained how they had arrived at their diagnosis. We analysed the data from the two test sessions to see if we could find elements of deliberate reflection in their explanations and whether we found differences in diagnostic accuracy between the conditions.

The second study, described in **Chapter 3**, tested learning-by-teaching as a means to teach deliberate reflection to general-practice residents, compared to a control condition. The study consisted of a learning session and a test session. In the learning session, par-

ticipants either studied examples of deliberate reflection and then explained the procedure and how it was applied to a fictitious peer while being video recorded, or they solved cases without deliberate reflection (control condition). In the test session, all participants took the same test in which they diagnosed new cases while thinking aloud. We looked for element of deliberate reflection in the think-aloud protocols and analysed diagnostic accuracy.

The third study, described in **Chapter 4**, tested whether medical students would learn deliberate reflection via learning-by-teaching, but would only apply it when they thought that cases would be difficult and therefore required extra attention. The study consisted of a learning phase and a test phase. The learning phase had the same instructions as the study in Chapter 3; Participants either studied examples of deliberate reflection and then explained the procedure and how it was applied to a fictitious peer while being video recorded, or they solved cases without deliberate reflection (control condition). In the test session, all participants took the same test in which they first diagnosed a case and then completed a recall task for the case, writing down everything they remembered, before moving on to the next case. When participants had seen half of the cases, they were told that the next set of cases would be difficult ones, although case difficulty did not actually change. The cases in the test phase were ambiguous cases that had two equally likely diagnoses. We analysed what students had recalled from the case to see whether they had mainly focussed on their own diagnosis, or on both diagnoses, which would be indicative of applying deliberate reflection.

The fourth study, described in **Chapter 5**, tested whether general-practice residents' diagnostic calibration could be improved with feedback. This study consisted of just one session in which residents diagnosed a set of cases. Participant in the feedback condition diagnosed a case, indicated how confident they were that their diagnosis was correct, and then got the correct diagnosis for the case to compare with their own diagnosis. Participants in the control condition followed the same procedure but without receiving feedback. We analysed whether diagnostic calibration differed between the two conditions.

**Chapter 6** provides a general discussion of the main results of the four studies, along with possible implications of the study results for medical education (and in particular, general practitioners in training) and ideas for directions for further research in this field.

## REFERENCES


Abrami, P. C., Bernard, R. M., Borokhovski, E., Waddington, D. I., Wade, C. A., & Persson, T. (2015). Strategies for Teaching Students to Think Critically:A Meta-Analysis. *Review of Educational Research*, *85*(2), 275-314. https://doi.org/10.3102/0034654314551063

Abrami, P. C., Bernard, R. M., Borokhovski, E., Wade, A., Surkes, M. A., Tamim, R., & Zhang, D. (2008). Instructional Interventions Affecting Critical Thinking Skills and Dispositions: A Stage 1 Meta-Analysis. *Review of Educational Research*, *78*(4), 1102-1134. https://doi.org/10.3102/0034654308326084

Atkinson, R. K., Derry, S. J., Renkl, A., & Wortham, D. (2000). Learning from Examples: Instructional Principles from the Worked Examples Research. *Review of Educational Research*, *70*(2), 181-214.

Berner, E. S., & Graber, M. L. (2008). Overconfidence as a Cause of Diagnostic Error in Medicine. *The American Journal of Medicine*, *121*(5 Supplement), S2-S23. https://doi.org/10.1016/j.amjmed.2008.01.001

Bhise, V., Rajan, S. S., Sittig, D. F., Vaghani, V., Morgan, R. O., Khanna, A., & Singh, H. (2018). Electronic health record reviews to measure diagnostic uncertainty in primary care. *Journal of Evaluation in Clinical Practice*, *24*(3), 545-551. https://doi.org/10.1111/jep.12912

Braun, L. T., Zwaan, L., Kiesewetter, J., Fischer, M. R., & Schmidmaier, R. (2017). Diagnostic errors by medical students: results of a prospective qualitative study. *BMC Medical Education*, *17*(1), 191.

Brown, A. L., & Kane, M. J. (1988). Preschool children can learn to transfer: Learning to learn and learning from example. *Cognitive Psychology*, *20*(4), 493-523. https://doi.org/https://doi.org/10.1016/0010-0285(88)90014-X

Chew, K. S., Durning, S. J., & van Merriënboer, J. J. (2016). Teaching metacognition in clinical decision-making using a novel mnemonic checklist: an exploratory study. *Singapore Medical Journal*, *57*(12), 694-700. https://doi.org/10.11622/smedj.2016015

Costa Filho, G. B., Moura, A. S., Brandão, P. R., Schmidt, H. G., & Mamede, S. (2019). Effects of deliberate reflection on diagnostic accuracy, confidence and diagnostic calibration in dermatology. *Perspectives on Medical Education*, *8*(4), 230-236. https://doi.org/10.1007/s40037-019-0522-5

Croskerry, P. (2003). The importance of cognitive errors in diagnosis and strategies to minimize them. *Academic Medicine*, *78*(8), 775 - 780.

Croskerry, P., Singhal, G., & Mamede, S. (2013). Cognitive debiasing 2: impediments to and strategies for change. *BMJ Quality & Safety*, *22 Suppl 2*(Suppl 2), ii65-ii72.

Davis, D. A., Mazmanian, P. E., Fordis, M., Van Harrison, R., Thorpe, K. E., & Perrier, L. (2006). Accuracy of Physician Self-assessment Compared With Observed Measures of CompetenceA Systematic Review. *JAMA*, *296*(9), 1094-1102. https://doi.org/10.1001/jama.296.9.1094

Duran, D. (2017). Learning-by-teaching. Evidence and implications as a pedagogical mechanism

[doi: 10.1080/14703297.2016.1156011]. *Innovations in Education and Teaching International*, *54*(5), 476-484. https://doi.org/10.1080/14703297.2016.1156011

Elstein, A. S., & Schwartz, A. (2002). Clinical problem solving and diagnostic decision making: selective review of the cognitive literature. *BMJ*, *324*(7339), 729-732. https://doi.org/10.1136/bmj.324.7339.729

Ely, J. W., Graber, M. L., & Croskerry, P. (2011). Checklists to reduce diagnostic errors. *Academic Medicine*, *86*(3), 307-313.

Eva, K. W., Neville, A. J., & Norman, G. R. (1998). Exploring the etiology of content specificity: factors influencing analogic transfer and problem solving. *Academic Medicine*, *73*(10 Suppl), S1-5.

Evans, J. S. B. T. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology*, *59*, 255-278.

Fiorella, L., & Mayer, R. E. (2013). The relative benefits of learning by teaching and teaching expectancy. *Contemporary Educational Psychology*, *38*(4), 281-288. https://doi.org/10.1016/j.cedpsych.2013.06.001

Fiorella, L., & Mayer, R. E. (2014). Role of expectations and explanations in learning by teaching. *Contemporary Educational Psychology*, *39*(2), 75-85. https://doi.org/10.1016/j.cedpsych.2014.01.001

Fiorella, L., & Mayer, R. E. (2016). Eight Ways to Promote Generative Learning. *Educational Psychology Review*, *28*(4), 717-741. https://doi.org/10.1007/s10648-015-9348-9

Friedman, C. P., Gatti, G. G., Franz, T. M., Murphy, G. C., Wolf, F. M., Heckerling, P. S., Fine, P. L., Miller, T. M., & Elstein, A. S. (2005). Do Physicians Know When Their Diagnoses Are Correct? Implications for Decision Support and Error Reduction. *Journal of General Internal Medicine*, *20*(4), 334-339. https://doi.org/10.1111/j.1525-1497.2005.30145.x

Gaal, S., Verstappen, W., Wolters, R., Lankveld, H., van Weel, C., & Wensing, M. (2011). Prevalence and consequences of patient safety incidents in general practice in the Netherlands: a retrospective medical record review study. *Implementation science : IS*, *6*, 37-37. https://doi.org/10.1186/1748-5908-6-37

Graber, M. L., Franklin, N., & Gordon, R. (2005). Diagnostic Error in Internal Medicine. *Archives of Internal Medicine*, *165*(13), 1493-1499. https://doi.org/10.1001/archinte.165.13.1493

Griffith, P. B., Doherty, C., Smeltzer, S. C., & Mariani, B. (2021). Education initiatives in cognitive debiasing to improve diagnostic accuracy in student providers: A scoping review. *Journal of the American Association of Nurse Practitioners*, *33*(11), 862-871. https://doi.org/10.1097/jxx.0000000000000479

Hacker, D. J., & Bol, L. (2019). Calibration and Self-Regulated Learning Making the Connections. In *The Cambridge Handbook of Cognition and Education* (pp. 647-677). Cambridge University Press. https://doi.org/10.1017/9781108235631.026

Higgs, J., & Jones, M. A. (2000). *Clinical reasoning in the health professions* (2nd ed. ed.). Butterworth-Heinemann.

Hoogerheide, V., Deijkers, L., Loyens, S. M., & Heijltjes, A. (2016). Gaining from explaining: Learning improves from explaining to fictitious others on video, not from writing to them. *Contemporary Educational Psychology*, *44*, 95-106. https://doi.org/http://dx.doi.org/10.1016/j.cedpsych.2016.02.005

Hoogerheide, V., Loyens, S. M. M., & Van Gog, T. (2014). Effects of creating video-based modeling examples on learning and transfer. *Learning and Instruction*, *33*, 108-119. https://doi.org/10.1016/j.learninstruc.2014.04.005

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. *Journal of Educational Psychology*, *111*(1), 45-56.

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. *Medical Education*, *48*, 796-805. https://doi.org/10.1111/medu.12435

Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux.

Kitsantas, A., Zimmerman, B. J., & Cleary, T. (2000). The Role of Observation and Emulation in the Development of Athletic Self-Regulation. *Journal of Educational Psychology*, *92*(4), 811-817.

Kuhn, G. J. (2002). Diagnostic errors. *Academic Emergency Medicine*, *9*(7), 740-750.

Labuhn, A. S., Zimmerman, B. J., & Hasselhorn, M. (2010). Enhancing students' self-regulation and mathematics performance: the influence of feedback and self-evaluative standards. *Metacognition and Learning*, *5*(2), 173-194. https://doi.org/10.1007/s11409-010-9056-2

Lachner, A., Hoogerheide, V., van Gog, T., & Renkl, A. (2021). Learning-by-Teaching Without Audience Presence or Interaction: When and Why Does it Work? *Educational Psychology Review*. https://doi.org/10.1007/s10648-021-09643-4

Lambe, K. A., Reilly, G., Kelly, B. D., & Curristan, S. (2016). Dual-process cognitive interventions to enhance diagnostic reasoning: a systematic review. *BMJ Quality & Safety*, *25*(10), 808. https://doi.org/10.1136/bmjqs-2015-004417

Lipko, A. R., Dunlosky, J., Hartwig, M. K., Rawson, K. A., Swan, K., & Cook, D. (2009). Using Standards to Improve Middle School Students' Accuracy at Evaluating the Quality of Their Recall. *Journal of Experimental Psychology: Applied*, *15*(4), 307-318.

Mamede, S., Schmidt, H., Rikers, R., Custers, E., Splinter, T., & Saase, J. (2010). Conscious thought beats deliberation without attention in diagnostic decision-making: At least when you are an expert. *Psychological Research*, *74*(6), 586-592.

Mamede, S., Schmidt, H. G., & Penaforte, J. C. (2008). Effects of reflective practice on the accuracy of medical diagnoses. *Medical Education*, *42*(5), 468-475. https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1365-2923.2008.03030.x

Mamede, S., Schmidt, H. G., Rikers, R. M., Penaforte, J. C., & Coelho-Filho, J. M. (2008). Influence of perceived difficulty of cases on physicians' diagnostic reasoning. *Academic Medicine*, *83*(12), 1210-1216. https://doi.org/10.1097/ACM.0b013e31818c71d7

Mamede, S., Schmidt, H. G., Rikers, R. M. J. P., Penaforte, J. C., & Coelho-Filho, J. M. (2007).

Breaking down automaticity: Case ambiguity and the shift to reflective approaches in clinical reasoning. *Medical Education*, *41*(12), 1185-1192. https://doi.org/10.1111/j.1365-2923.2007.02921.x

Mamede, S., Splinter, T. A., Van Gog, T., Rikers, R. M., & Schmidt, H. G. (2012). Exploring the role of salient distracting clinical features in the emergence of diagnostic errors and the mechanisms through which reflection counteracts mistakes. *BMJ Quality & Safety*, *21*(4), 295-300. https://doi.org/10.1136/bmjqs-2011-000518

Mamede, S., Van Gog, T., Van den Berge, K., Rikers, R. M., Van Saase, J. L., Van Guldener, C., & Schmidt, H. G. (2010). Effect of Availability Bias and Reflective Reasoning on Diagnostic Accuracy Among InternalMedicine Residents. *JAMA*, *304*(11), 1198-1203. https://doi.org/10.1001/jama.2010.1276

Mendel, R., Traut-Mattausch, E., Jonas, E., Leucht, S., Kane, J. M., Maino, K., Kissling, W., & Hamann, J. (2011). Confirmation bias: why psychiatrists stick to wrong preliminary diagnoses. *Psychological Medicine*, *41*(12), 2651-2659. https://doi.org/Doi: 10.1017/s0033291711000808

Meyer, A. N. D., Payne, V. L., Meeks, D. W., Rao, R., & Singh, H. (2013). Physicians' diagnostic accuracy, confidence, and resource requests: a vignette study. *JAMA Internal Medicine*, *173*(21), 1952-1958. https://doi.org/10.1001/jamainternmed.2013.10081

Meyer, A. N. D., & Singh, H. (2019). The Path to Diagnostic Excellence Includes Feedback to Calibrate How Clinicians Think. *JAMA*, *321*(8), 737-738. https://doi.org/10.1001/jama.2019.0113

Monteiro, S., Sherbino, J., Patel, A., Mazzetti, I., Norman, G. R., & Howey, E. (2015). Reflecting on Diagnostic Errors: Taking a Second Look is Not Enough. *Journal of General Internal Medicine*, *30*(9), 1270-1274. https://doi.org/10.1007/s11606-015-3369-4

Monteiro, S., Sherbino, J., Sibbald, M., & Norman, G. (2020). Critical thinking, biases and dual processing: The enduring myth of generalisable skills. *Medical Education*, *54*(1), 66-73. https://doi.org/https://doi.org/10.1111/medu.13872

Nederhand, M. L., Tabbers, H. K., & Rikers, R. M. J. P. (2019). Learning to calibrate: Providing standards to improve calibration accuracy for different performance levels. *Applied Cognitive Psychology*, *33*(6), 1068-1079. https://doi.org/10.1002/acp.3548

Nederhand, M. L., Tabbers, H. K., Splinter, T. A. W., & Rikers, R. M. J. P. (2018). The Effect of Performance Standards and Medical Experience on Diagnostic Calibration Accuracy. *Health Professions Education*, *4*(4), 300-307. https://doi.org/10.1016/j.hpe.2017.12.008

Norman, G. R., Monteiro, S., Sherbino, J., Ilgen, J. S., Schmidt, H. G., & Mamede, S. (2017). The Causes of Errors in Clinical Reasoning: Cognitive Biases, Knowledge Deficits, and Dual Process Thinking. *Academic Medicine*, *92*(1), 23-30. https://doi.org/10.1097/acm.0000000000001421

Paas, F., & Van Gog, T. (2006). Optimising worked example instruction: Different ways to increase germane cognitive load. *Learning and Instruction*, *16*(2), 87-91. https://doi.org/https://doi.org/10.1016/j.learninstruc.2006.02.004

Patel, V. L., Kaufman, D. R., & Kannampallil, T. G. (2013). Diagnostic Reasoning and Decision Making in the Context of Health Information Technology. *Reviews of Human Factors and Ergo-*

*nomics*, *8*(1), 149-190. https://doi.org/10.1177/1557234x13492978

Prakash, S., Sladek, R. M., & Schuwirth, L. (2019). Interventions to improve diagnostic decision making: A systematic review and meta-analysis on reflective strategies. *Medical Teacher*, *41*(5), 517-524. https://doi.org/10.1080/0142159x.2018.1497786

Rummel, N., & Spada, H. (2005). Learning to Collaborate: An Instructional Approach to Promoting Collaborative Problem Solving in Computer-Mediated Settings. *Journal of the Learning Sciences*, *14*(2), 201-241.

Schmidt, H. G., & Boshuizen, H. P. A. (1993). On Acquiring Expertise in Medicine. *Educational Psychology Review*, *5*(3), 205-221.

Schmidt, H. G., & Mamede, S. I. (2015). How to improve the teaching of clinical reasoning: a narrative review and a proposal. *Medical Education*, *49*(10), 961-973. https://doi.org/10.1111/medu.12775

Schmidt, H. G., Mamede, S. I., Van Den Berge, K., Van Gog, T., Van Saase, J. L. C. M., & Rikers, R. M. J. P. (2014). Exposure to media information about a disease can cause doctors to misdiagnose similar-looking clinical cases. *Academic Medicine*, *89*(2).

Schmidt, H. G., Norman, G., & Boshuizen, H. (1990). A cognitive perspective on medical expertise: theory and implications. *Academic Medicine*, *65*(10), 611-621.

Schmidt, H. G., & Rikers, R. M. J. P. (2007). How expertise develops in medicine: knowledge encapsulation and illness script formation. *Medical Education*, *41*(12), 1133-1139. https://doi.org/10.1111/j.1365-2923.2007.02915.x

Schmidt, H. G., Van Gog, T., Schuit, S. C. E., Van Den Berge, K., Van Daele, P. L. A., Bueving, H., Van der Zee, T., Van Den Broek, W. W., Van Saase, J. L. C. M., & Mamede, S. I. (2017). Do patients' disruptive behaviours influence the accuracy of a doctor's diagnosis? A randomised experiment. *BMJ Quality & Safety*, *26*(1), v19-23. https://doi.org/10.1136/bmjqs-2015-004109

Sherbino, J., Kulasegaram, K., Howey, E., & Norman, G. (2014). Ineffectiveness of cognitive forcing strategies to reduce biases in diagnostic reasoning: a controlled trial. *Cjem*, *16*(1), 34-40.

Shimizu, T., Matsumoto, K., & Tokuda, Y. (2013). Effects of the use of differential diagnosis checklist and general de-biasing checklist on diagnostic performance in comparison to intuitive diagnosis. *Medical Teacher*, *35*(6), e1218-1229.

Sibbald, M., & de Bruin, A. B. (2012). Feasibility of self-reflection as a tool to balance clinical reasoning strategies. *Adv Health Sci Educ Theory Pract*, *17*(3), 419-429.

Stark, R., Kopp, V., & Fischer, M. R. (2011). Case-based learning with worked examples in complex domains: Two experimental studies in undergraduate medical education. *Learning and Instruction*, *21*(1), 22-33. https://doi.org/10.1016/j.learninstruc.2009.10.001

Van Gog, T., & Rummel, N. (2010). Example-Based Learning: Integrating Cognitive and Social-Cognitive Research Perspectives. *Educational Psychology Review*, *22*(2), 155-174. https://doi.org/10.1007/s10648-010-9134-7

Van Gog, T., Rummel, N., & Renkl, A. (2019). Learning How to Solve Problems by Studying Examples. In J. Dunlosky & K. A. Rawson (Eds.), *The Cambridge Handbook of Cognition and Education*

(pp. 183-208). Cambridge University Press. https://doi.org/Doi: 10.1017/9781108235631.009

Van Merriënboer, J. J., & Sweller, J. (2010). Cognitive load theory in health professional education: design principles and strategies. *Medical Education*, *44*(1), 85-93.

Voytovich, A. E., Rippey, R. M., & Suffredini, A. (1985). Premature conclusions in diagnostic reasoning. *Journal of Medical Education*, *60*(4), 302-307.

Zwaan, L., Thijs, A., Wagner, C., & Timmermans, D. R. (2013). Does inappropriate selectivity in information use relate to diagnostic errors and patient harm? The diagnosis of patients with dyspnea. *Social Science and Medicine*, *91*, 32-38.

Zwaan, L., Thijs, A., Wagner, C., van der Wal, G., & Timmermans, D. R. (2012). Relating faults in diagnostic reasoning with diagnostic errors and patient harm. *Academic Medicine*, *87*(2), 149-156.

2

# CHAPTER 2

# Can we Teach Reflective Reasoning in General-Practice Training through Example-based Learning and Learning-by-doing?

**This chapter has been published as:**

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 Professions Education, 6*(4), 506-515. https://doi.org/10.1016/j.hpe.2020.07.004

## 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.

## 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 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 (Mamede et al., 2008). 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 (Mamede et al., 2017; Mamede, Van Gog, Van den Berge, et al., 2014). If a wrong initial diagnostic hypothesis has been generated, the mistake could only be corrected by further reflection on the case (Hess et al., 2015).

Studies on deliberate reflection have found that it can counteract diagnostic errors on complex cases (Mamede et al., 2008), or if physicians were distracted by irrelevant patient features (Mamede, Splinter, et al., 2012; Schmidt et al., 2017) or influenced by other cases they had encountered recently (i.e. availability bias) (Mamede, Splinter, et al., 2012; Mamede et al., 2010). Deliberate reflection has been investigated as a learning tool as well. Students ($4^{th}$ – $6^{th}$ 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 (Ibiapina et al., 2014; Mamede, Van Gog, et al., 2012; Mamede, Van Gog, Moura, et al., 2014).

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 (Atkinson et al., 2000; Van Gog & Rummel, 2010). 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 (Sweller, 1988). 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 (Bjerrum et al., 2013) and diagnostic competence (Stark et al., 2011). It can therefore be hypothesised that EBL would be effective to teach deliberate reflection as well, if learners have never worked with it before.

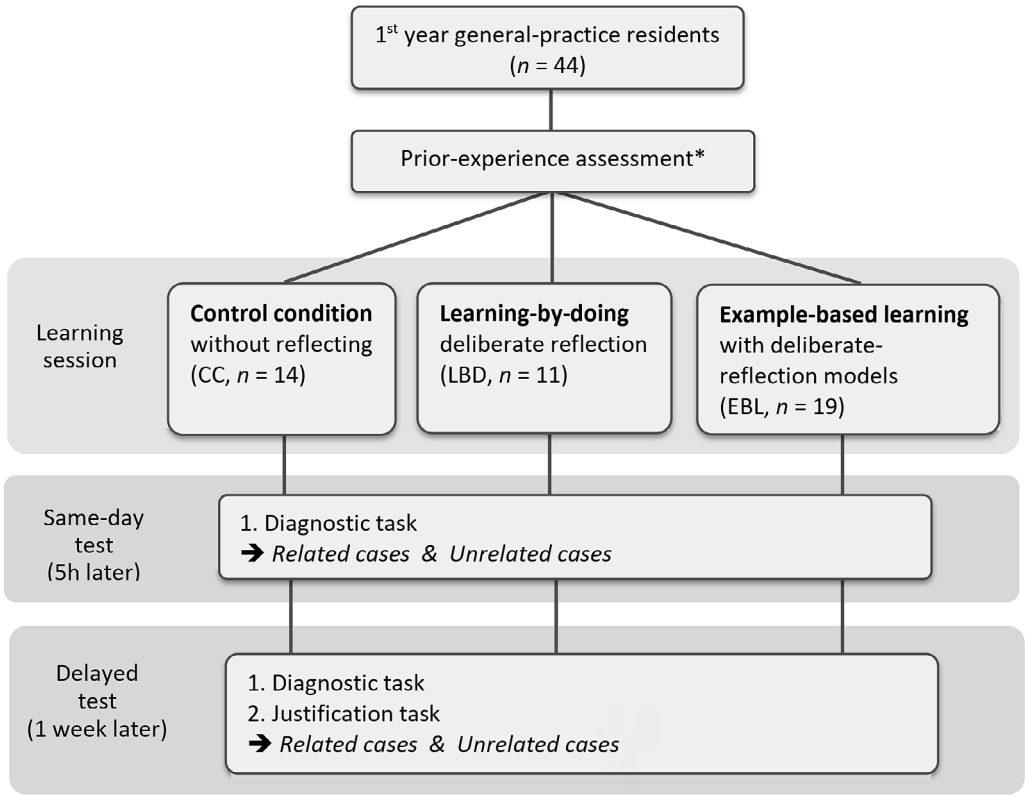
In this study, we investigated whether the deliberate-reflection procedure can be learned and then be applied autonomously on future cases, and which teaching approach is most effective for residents in general-practice training. For this purpose, we conducted an experiment consisting of a learning session, 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).

# METHOD

## Design

The study consisted of a prior-experience assessment and three sessions (Figure 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.

Figure 1 - Illustration of the study protocol.




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

## 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 $f = 0.25$) (Cohen, 1988) at $\alpha = 0.05$, showed that a sample of 81 would be sufficient to have a power of 0.80. The study took place during the usual educational program, and participants did not receive compensation.

## Material and procedure

All material was presented in Dutch. Thirty written cases were used in this study (Appendix A), 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 Appendix B) 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.

**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 (Appendix C). 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).

**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.



***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 (Paas, 1992). 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.

***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 (Mamede et al., 2008). 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.

***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 Appendix D). 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.

**Same-day test session.** The same-day test was conducted three to five hours 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.

**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 (Appendix E), 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.

## 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, $\eta_p^2$ 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 (Cohen, 1988).

**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.

**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, (Cicchetti, 1994). 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 (Meyer, 1975; Schiefele & Krapp, 1996) that could be categorised according to the deliberate-reflection steps 1 – 4 (example in Appendix F). 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 (Cicchetti, 1994) (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 (Mamede & Schmidt, 2014). 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-subjects factors, and study condition (EBL, LBD, control) as a between-subjects factor.

## RESULTS

### 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 G). 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.

### Prior experience

The response rate on the prior-experience questionnaire was 79.54% (means in Appendix G). 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 (Appendix H).

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 | $F(2, 41) = 2.33$, $p = .110$, $\eta_p^2 = .10$ | $F(1, 41) = 1.59$, $p = .214$, $\eta_p^2 = .04$ | $F(2, 41) = .15$, $p = .855$, $\eta_p^2 = .08$ |

## Same-day test

Means and standard deviations are shown in Table 2. 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, $p = .099$.

Table 2 - 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).

## Delayed test

**Diagnostic task.** Means and standard deviations are shown in Table 3. 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, $p = .544$.

**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, $p = .624$.

Table 3 - 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).

## 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 (Hoogerheide et al., 2014). 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 (Atkinson et al., 2000; Van Gog & Rummel, 2010), did not apply to them. In studies where residents' diagnostic accuracy was improved by deliberate reflection (Mamede et al., 2008; Mamede, Splinter, et al., 2012; Schmidt et al., 2017), 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 (Norman et al., 2017; Schmidt & Mamede, 2015). 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 (Ibiapina et al., 2014; Mamede, Van Gog, et al., 2012; Mamede, Van Gog, Moura, et al., 2014), 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) (Roediger & Karpicke, 2006). 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.

## REFERENCES

Atkinson, R. K., Derry, S. J., Renkl, A., & Wortham, D. (2000). Learning from Examples: Instructional Principles from the Worked Examples Research. *Review of Educational Research*, *70*(2), 181-214.

Bjerrum, A. S., Hilberg, O., Van Gog, T., Charles, P., & Eika, B. (2013). Effects of modelling examples in complex procedural skills training: a randomised study. *Medical Education*, *47*(9), 888-898. https://doi.org/10.1111/medu.12199

Cicchetti, D. V. (1994). Guidelines, Criteria, and Rules of Thumb for Evaluating Normed and Standardized Assessment Instruments in Psychology. *Psychological Assessment*, *6*(4), 284-290.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2 ed.). Lawrence Erlbaum Associates.

Hess, B. J., Lipner, R. S., Thompson, V., Holmboe, E. S., & Graber, M. L. (2015). Blink or think: can further reflection improve initial diagnostic impressions? *Academic Medicine*, *90*(1), 112-118. https://doi.org/10.1097/acm.0000000000000550

Hoogerheide, V., Loyens, S. M. M., & Van Gog, T. (2014). Effects of creating video-based modeling examples on learning and transfer. *Learning and Instruction*, *33*, 108-119. https://doi.org/10.1016/j.learninstruc.2014.04.005

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. *Medical Education*, *48*, 796-805. https://doi.org/10.1111/medu.12435

Mamede, S., & Schmidt, H. G. (2014). Reflection in Diagnostic Reasoning: What Really Matters? *Academic Medicine*, *89*(7), 959-960. https://doi.org/10.1097/acm.0000000000000306

Mamede, S., Schmidt, H. G., & Penaforte, J. C. (2008). Effects of reflective practice on the accuracy of medical diagnoses. *Medical Education*, *42*(5), 468-475. https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1365-2923.2008.03030.x

Mamede, S., Splinter, T. A., Van Gog, T., Rikers, R. M., & Schmidt, H. G. (2012). Exploring the role of salient distracting clinical features in the emergence of diagnostic errors and the mechanisms through which reflection counteracts mistakes. *BMJ Quality & Safety*, *21*(4), 295-300. https://doi.org/10.1136/bmjqs-2011-000518

Mamede, S., Van Gog, T., Moura, A. S., De Faria, R. M., Peixoto, J. M., Rikers, R. M., & Schmidt, H. G. (2012). Reflection as a strategy to foster medical students' acquisition of diagnostic competence. *Medical Education*, *46*(5), 464-472. https://doi.org/10.1111/j.1365-2923.2012.04217.x

Mamede, S., Van Gog, T., Moura, A. S., de Faria, R. M. D., Peixoto, J. M., & Schmidt, H. G. (2014). 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. *Academic Medicine*, *89*, 121-127.

Mamede, S., Van Gog, T., Schuit, S. C., Van den Berge, K., Van Daele, P. L., Bueving, H., Van

der Zee, T., Van den Broek, W. W., Van Saase, J. L., & Schmidt, H. G. (2017). Why patients' disruptive behaviours impair diagnostic reasoning: a randomised experiment. *BMJ Quality & Safety*, *26*(1), 13-18. https://doi.org/10.1136/bmjqs-2015-005065

Mamede, S., Van Gog, T., Van den Berge, K., Rikers, R. M., Van Saase, J. L., Van Guldener, C., & Schmidt, H. G. (2010). Effect of Availability Bias and Reflective Reasoning on Diagnostic Accuracy Among InternalMedicine Residents. *JAMA*, *304*(11), 1198-1203. https://doi.org/10.1001/jama.2010.1276

Mamede, S., Van Gog, T., Van den Berge, K., Van Saase, J. L. C. M., & Schmidt, H. G. (2014). Why Do Doctors Make Mistakes? A Study of the Role of Salient Distracting Clinical Features. *Academic Medicine*, *89*(1), 114-120. https://doi.org/10.1097/acm.0000000000000077

Meyer, B. J. F. (1975). *The organization of prose and its effects on memory*. Amsterdam: North-Holland Publishing Co.

New York: American Elsevier Publishing Co.

Norman, G. R., Monteiro, S., Sherbino, J., Ilgen, J. S., Schmidt, H. G., & Mamede, S. (2017). The Causes of Errors in Clinical Reasoning: Cognitive Biases, Knowledge Deficits, and Dual Process Thinking. *Academic Medicine*, *92*(1), 23-30. https://doi.org/10.1097/acm.0000000000001421

Paas, F. (1992). Training Strategies for Attaining Transfer of Problem-Solving Skill in Statistics: A Cognitive-Load Approach. *Journal of Educational Psychology*, *84*(4), 429-434. https://doi.org/10.1037/0022-0663.84.4.429

Roediger, H. L., & Karpicke, J. D. (2006). The Power of Testing Memory: Basic Research and Implications for Educational Practice. *Perspectives on Psychological Science*, *1*(3), 181-210.

Schiefele, U., & Krapp, A. (1996). Topic interest and free recall of expository text [doi:10.1016/S1041-6080(96)90030-8]. *Learning and Individual Differences*, *8*(2), 141-160. https://doi.org/10.1016/s1041-6080(96)90030-8

Schmidt, H. G., & Mamede, S. I. (2015). How to improve the teaching of clinical reasoning: a narrative review and a proposal. *Medical Education*, *49*(10), 961-973. https://doi.org/10.1111/medu.12775

Schmidt, H. G., Van Gog, T., Schuit, S. C. E., Van Den Berge, K., Van Daele, P. L. A., Bueving, H., Van der Zee, T., Van Den Broek, W. W., Van Saase, J. L. C. M., & Mamede, S. I. (2017). Do patients' disruptive behaviours influence the accuracy of a doctor's diagnosis? A randomised experiment. *BMJ Quality & Safety*, *26*(1), v19-23. https://doi.org/10.1136/bmjqs-2015-004109

Stark, R., Kopp, V., & Fischer, M. R. (2011). Case-based learning with worked examples in complex domains: Two experimental studies in undergraduate medical education. *Learning and Instruction*, *21*(1), 22-33. https://doi.org/10.1016/j.learninstruc.2009.10.001

Sweller, J. (1988). Cognitive Load During Problem Solving: Effects on Learning. *Cognitive Science*, *12*(2), 257-285. https://doi.org/https://doi.org/10.1016/0364-0213(88)90023-7

Van Gog, T., & Rummel, N. (2010). Example-Based Learning: Integrating Cognitive and Social-Cognitive Research Perspectives. *Educational Psychology Review*, *22*(2), 155-174. https://doi.org/10.1007/s10648-010-9134-7

# APPENDIX A

*Overview of the medical conditions of the cases used during the different sessions. Chief symptoms of unrelated cases are given in parenthesis.*

| Learning session | Same-day test | Delayed test |
|---|---|---|
| Studied cases of diarrhoea | Related cases of diarrhoea | Related cases of diarrhoea |
| Diverticulitis | Diverticulitis | Diverticulitis |
| Irritable bowel syndrome | Irritable bowel syndrome | Irritable bowel syndrome |
| | Chronic pancreatitis | Chronic pancreatitis |
| | Inflammatory bowel disease | Inflammatory bowel disease |
| Studied cases of shortness of breath | Related cases of shortness of breath | Related cases of shortness of breath |
| Infection of the lower respiratory tract | Infection of the lower respiratory tract | Infection of the lower respiratory tract |
| Heart failure | Heart failure | Heart failure |
| | Asthma | Asthma |
| | Pulmonary embolism | Pulmonary embolism |
| Studied cases of various chief symptoms | Unrelated cases of various novel chief symptoms | Unrelated cases of various novel chief symptoms |
| Pregnancy (amenorrhea) | Panic disorder (palpitation) | Benign Paroxysmal Position Vertigo (turn dizziness) |
| Multiple sclerosis (tremor in hand) | Scarlet fever (rash / eczema) | Bell's palsy (facial paralysis) |
| | Spondylodiscitis (lower back pain) | Rosacea (Rash in the face) |
| | Spinal canal stenosis (pain in legs) | Bacterial vaginosis (vaginal discharge) |

# APPENDIX B

*Translation of a Dutch case of diverticulitis, with as chief complaint diarrhoea. This case has been shown during the learning session.*

You receive the medical record for Mrs. de Vries (73 years) from your assistant. She called because she has had diarrhoea twice since this morning. She also has a lot of pain on the left side of her abdomen. You wonder why you have to go here, because it makes you think of a gastroenteritis. As you arrive at Mrs. de Vries's house her worried partner guides you to her bedroom.

Mrs. de Vries lies in bed and indeed has a lot of pain, which feels sharp. In addition, she has loose stools. Yesterday there was nothing wrong. It was a busy day. They just had the terrace refurbished and yesterday it was very nice summer weather. So the children have been along and they all have barbecued together in the garden. She has not had this pain before. You can see in her medical record that she gets Methotrexate for her rheumatoid arthritis. Besides that there is nothing special in her record. You ask Mrs. de Vries some questions for an anamnesis of her gastrointestinal tract, and it turns out that she did not have to vomit. However, she does have less appetite.

When you wash your hands after the physical examination, her partner, who you do not know yet, addresses you in the bathroom. He worries whether this is all happening because of the barbeque. After all, it was he who prepared the chicken legs. The barbeque was the first occasion on which he got to meet the children of Mrs. de Vries.

Physical examination:
Saturation 97%, pulse 86, blood pressure 156/94, temperature 38.2 °C. abdomen: sparse peristalsis, alternating tympanum, pressure and release pain in left lower abdomen

# APPENDIX C

*Diagnoses and symptoms shown during the prior-experience assessment.*

| Correct diagnoses and chief symptoms that appear in the cases | Other diagnoses and symptoms (filler) |
|---|---|
| Diarrhoea | Constipation |
| Diverticulitis | Abdominal pain |
| Irritable bowel syndrome (IBS) | Gastroenteritis |
| Inflammatory bowel disease (IBD) | Infection of the upper respiratory tract |
| Chronic pancreatitis | Chronic Obstructive Pulmonary Disease |
| Shortness of breath | Anaemia |
| Infection of the lower respiratory tract | Pain on the chest |
| Asthma | Depression |
| Pulmonary embolism | Eating disorder |
| Heart failure | Vaginal complaints |
| Palpitations | Vaginal fungal infection |
| Anxiety / panic disorder | Sexually Transmitted Diseases |
| Vaginal discharge | Erectile dysfunction |
| Amenorrhea | Headache |
| Bacterial vaginosis | Dizziness |
| Pregnancy | Thumb base instability |
| Turn dizziness | Shingles |
| Benign Paroxysmal Position Vertigo (BPPD) | Acne |
| Skin rash | Pain in thumb |
| Rash in the face | Weak muscles |
| Scarlet fever | Hyperthyroidism |
| Rosacea | Cerebrovascular Accident (CVA) |
| Lower back pain | Lyme disease |
| Pain in legs | |
| Spondylodiscitis | |
| Spinal canal stenosis / neurogenic claudication | |
| Tremor in hand | |
| Multiple sclerosis | |
| Facial paralysis | |
| Idiopathic peripheral facial paralysis (IPAV) | |

# APPENDIX D

*Translation of the Dutch deliberate-reflection model shown to participants in the example-based-learning condition during the learning session. Participants in the deliberate-reflection condition received a similar empty table, which showed only the instruction in the first row. These participants filled in the diagnoses and clinical findings themselves.*

| Diagnosis | Which findings from the case speak **for** this diagnosis? | Which findings from the case speak **against** this diagnosis? | Which further findings would you expect if your diagnosis was correct, which are **missing** for this patient? |
|---|---|---|---|
| Gastroenteritis | - Diarrhoea<br>- Possible source (barbeque)<br>- Increased susceptibility due to immuno-suppressants<br>- Temperature 38.2 °C | - Pain only on left lower side | - Nausea & vomiting<br>- diffuse, cramping abdominal pain<br>- Others who have eaten from the barbeque should have complaints |
| Diverticulitis (complicated/ uncomplicated) | - Character of the pain: sharp & stabbing, bottom left<br>- Pressure & release pain only left lower side<br>- Temperature 38.2 °C<br>- No alarm signals<br>- short duration of complaint (one day) | | |
| Colorectal carcinoma | - Diarrhoea<br>- Pressure and release pain in left lower abdomen<br>- Age | - Acute start<br>- Short duration of complaints<br>- Temperature 38.2 °C | - Rectal blood loss<br>- Weight loss<br>- Resistance in the abdomen<br>- Family anamnesis |

## APPENDIX E

*Translation of a Dutch case that has been shortened for the justification task. The text describes some features, including the chief complaint, so that participants can recognise the case.*

Ms Duinkerken (50 years) (finally) got divorced about 2 years ago, after a relationship in which there was domestic violence. You suspect that, during that period, she drank more alcohol than was good for her. (...)
You startle when you call her into the consulting room. She got pretty emaciated since you saw her the last time! She says that she suffers from diarrhoea, about 3 - 4 times a day. She also has increasing pain in the upper abdomen since ½ year.



## APPENDIX F

*Translation of one participant's answer on the justification task during the delayed test and how we counted the idea units in that answer, categorized into the deliberate reflection steps*

Participant's justification: "Complaints sounded like asthma, but that does not fit with the leukocyte differentiation. Lung noises and slightly increased CRP and fever?? That reminded me of bronchitis"

| *Diagnoses* | *Which findings from the case speak for this diagnosis?* | *Which findings from the case speak against this diagnosis?* | *Which further findings would you expect if your diagnosis was correct, which are missing for this patient?* |
|---|---|---|---|
| *Initial diagnosis:* Asthma | - Complaints sounded like asthma (1) | - but that does not fit with the leukocyte differentiation (1) | (0) |
| Bronchitis (1) | - Lung noises<br>- slightly increased CRP<br>- and fever (3) | (0) | (0) |

## APPENDIX G

*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 H

*Prior experience in medical practice.*

| | *Frequency* | *Percentage* |
|---|---|---|
| **Did his/her main clerkship in general practice** | | |
| Control (*n* = 12) | 8 | 66.67 |
| EBL (*n* = 12) | 5 | 41.67 |
| LBD (*n* = 11) | 0 | 0.00 |
| Total (*N* = 35) | 9 | 25.71 |
| **Has worked as a senior house officer in general practice** | | |
| Control (*n* = 12) | 0 | 0.00 |
| EBL (*n* = 12) | 1 | 8.33 |
| LBD (*n* = 11) | 1 | 9.01 |
| Total (*N* = 35) | 2 | 5.71 |
| **Has worked as a senior house officer in another specialisation** | | |
| Control (*n* = 12) | 11 | 91.67 |
| EBL (*n* = 12) | 12 | 100.00 |
| LBD (*n* = 11) | 10 | 90.91 |
| Total (*N* = 35) | 33 | 94.29 |

*Note.* Only 35 of the 44 participants filled in the prior-experience questionnaire (79.54%). With senior house officer, we mean a postgraduate physician who is not (yet) in specialty training (in Dutch called *ANIOS*).

# CHAPTER 3

# Learning Deliberate Reflection in Medical Diagnosis: Does Learning-by-Teaching Help?

**This chapter has been published as:**

Kuhn, J., Mamede, S., van den Berg, P., Zwaan, L., van Peet, P., Bindels, P., & van Gog, T. (2022). Learning deliberate reflection in medical diagnosis: does learning-by-teaching help? *Advances in Health Sciences Education*. https://doi.org/10.1007/s10459-022-10138-2

## ABSTRACT

Deliberate reflection has been found to foster diagnostic accuracy on complex cases or under circumstances that tend to induce cognitive bias. However, it is unclear whether the procedure can also be learned and thereby autonomously applied when diagnosing future cases without instructions to reflect. We investigated whether general practice residents would learn the deliberate reflection procedure through 'learning-by-teaching' and apply it to diagnose new cases. The study was a two-phase experiment. In the learning phase, 56 general-practice residents were randomly assigned to one of two conditions. They either (1) studied examples of deliberate reflection and then explained the procedure to a fictitious peer on video; or (2) solved cases without reflection (control). In the test phase, one to three weeks later, all participants diagnosed new cases while thinking aloud. The analysis of the test phase showed no significant differences between the conditions on any of the outcome measures (diagnostic accuracy, $p$ = . 263; time to diagnose, $p$ = .598; mental effort ratings, $p$ = .544; confidence ratings, $p$ = .710; proportion of contradiction units (i.e. measure of deliberate reflection), $p$ = .544). In contrast to findings on learning-by-teaching from other domains, teaching deliberate reflection to a fictitious peer, did not increase reflective reasoning when diagnosing future cases. Potential explanations that future research might address are that either residents in the experimental condition did not apply the learned deliberate reflection procedure in the test phase, or residents in the control condition also engaged in reflection.

## INTRODUCTION

Reflection upon one's own experiences has been much valued as a means for physicians to learn and improve performance throughout their professional life (Mann et al., 2009; Ng et al., 2015). Reflection may have different foci, occur at different moments of practice, and there are many ways for physicians to engage in reflection (Ng et al., 2015). While reflection in a broader sense can be seen as the ability to critically examine one's own explanation for or beliefs about a problem (Dewey, 1910), the *deliberate reflection* (Mamede, Schmidt, & Penaforte, 2008) procedure has been developed to facilitate structured reflection and avoid biased reasoning on to-be-diagnosed clinical cases. In two recent reviews about the effectiveness of cognitive interventions to improve diagnostic accuracy, this procedure showed to be among the most effective and consistently successful interventions to improve diagnostic accuracy (Lambe et al., 2016; Prakash et al., 2019). Deliberate reflection consists of specific instructions for stepwise consideration of initial diagnostic hypothesis and alternative diagnoses. Physicians first read the case and give an initial diagnosis. After that, they are asked to list all the findings that speak for and against their initial diagnosis for the case, as well as findings that they would expect with their diagnosis, which are absent. Then they are asked to generate alternative diagnoses and do the same 'reflective steps' for those. When a couple of diagnoses have been analysed, they rank the diagnoses in order of likelihood to make a decision on their final diagnosis. This procedure aims to stimulate physicians to reflect on their first impression of a case to avoid excessive reliance on intuitive reasoning.



Deliberate reflection has been shown to improve diagnostic accuracy, especially when cases are complex (Mamede, Schmidt, et al., 2010; Mamede, Schmidt, & Penaforte, 2008), or when physicians diagnose cases under conditions that tend to induce cognitive biases that mislead diagnostic reasoning (Mamede, Van Gog, et al., 2010; Schmidt et al., 2014; Schmidt et al., 2017). For example, when they have just recently seen a case that resembles the one at hand on superficial features, deliberate reflection helps physicians not to be misled by these similarities into thinking that they have the same clinical condition when they do not, or when patients show disruptive behaviour, deliberate reflection can help physicians to better focus on the clinical findings and avoid diagnostic error. These studies have mainly focussed on a direct improvement in performance (i.e., diagnostic accuracy on the case reflected upon). However, it is as yet unclear whether the procedure itself can be learned and would then be applied autonomously (i.e., without reflection instructions) when diagnosing future cases. It has been questioned whether reasoning processes can be taught at all, as physicians engage in it unconsciously and interventions to teach diagnostic reasoning have often been found to be ineffective in improving diagnostic accuracy (Norman et al., 2017; Schmidt & Mamede, 2015). On the other hand, literature on example-based learning shows

that specific procedures can be learned and applied to new problems, and that this does not only apply to cognitive skills, for example in physics (Hoogerheide, Renkl, et al., 2019) or mathematics (Paas, 1992), but also to higher order skills such as collaboration (Rummel & Spada, 2005). Therefore, similar interventions may be useful to teach the steps of the deliberate-reflection procedure.

Example-based learning has proven very effective and efficient for many types of cognitive and higher order skills (Atkinson et al., 2000). However, in a previous study, the attempt to teach deliberate reflection by studying examples of experts' reflection on cases proved to be ineffective (Kuhn et al., 2020). Perhaps studying the examples was not sufficiently challenging. In order to learn a new problem-solving procedure and be able to transfer it to novel problems, students should actively engage with the study material (Brown & Kane, 1988; Eva et al., 1998). Learning-by-teaching could improve the effectiveness of example-based learning as it stimulates such active engagement. The present study investigated whether 'learning-by-teaching' (Fiorella & Mayer, 2013, 2014; Hoogerheide et al., 2016; Hoogerheide et al., 2014), an instructional method that has proven effective for enhancing learning and transfer to novel contexts, would be an effective way to learn (to adopt) the procedure.

Previous studies have found, that when students study material with the expectation to teach, this alone can have a short-term benefit on learning (Fiorella & Mayer, 2013, 2014). When students then also teach the material, they seem to develop a deeper understanding of the material and a benefit on learning is found even after a one-week delay. Explaining study material helps students to actively process it and to understand its important aspects and underlying rationale (Van Gog & Rummel, 2010). This helps with learning of a new problem-solving procedure and with applying it to slightly novel problems (i.e., transfer). Another benefit of learning-by-teaching is that students practice to retrieve the material from memory while teaching (Koh et al., 2018) which improves long term retention of the material (see testing effect; Roediger & Karpicke, 2006). Some studies have included measures of perceived mental effort because in combination with performance it can help to investigate the efficiency of the instructional method (Van Gog & Paas, 2008). These studies found, that learning-by-teaching is typically more cognitively demanding than restudying the material (i.e., participants usually perceive teaching as being more effortful), but this additional effort pays off, as they show better learning results (Hoogerheide et al., 2016; Hoogerheide, Renkl, et al., 2019; Hoogerheide, Visee, et al., 2019). Furthermore, it has been found to be more effective if students have a (perceived) audience (in the form of a camera), than when they teach without audience. The reason for this may be, that this feeling of a *social presence* of an audience increases active processing of the material (Hoogerheide et al., 2016) and arousal (Hoogerheide, Renkl, et al., 2019), which can foster learning.

We build on a recent study (Hoogerheide, Renkl, et al., 2019), in which psychology students were taught how to solve physics problems through 'learning-by-teaching': they first studied an example and then recorded a video on which they explained to a fictitious peer how to solve the problem. On a post-test, students who engaged in learning-by-teaching outperformed students who studied an additional example instead.

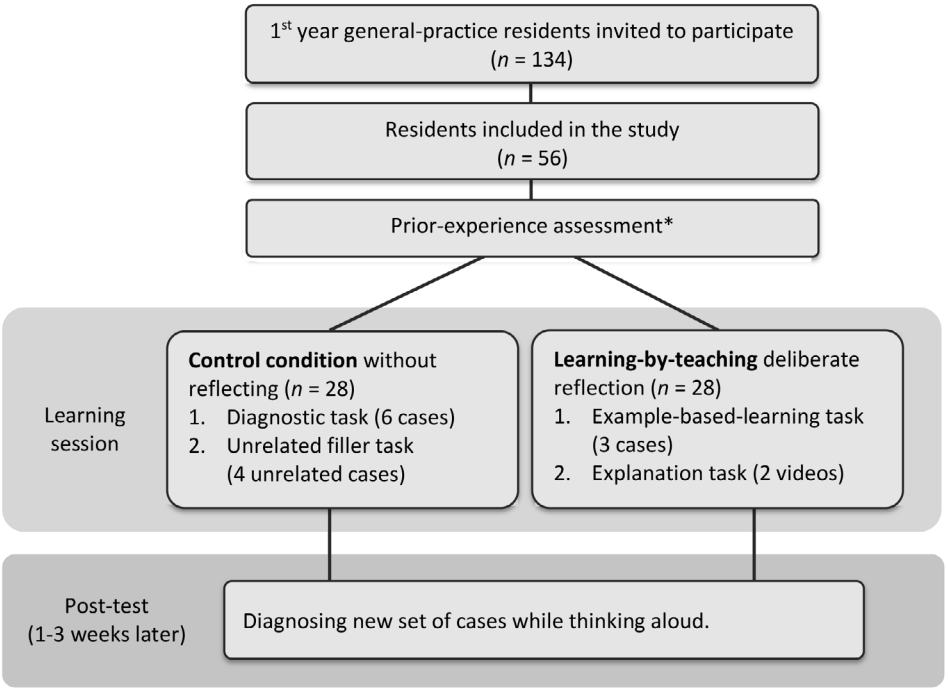
In the present study, we investigated whether general practice residents, i.e. physicians in training to become specialists, would learn the method of deliberate reflection by studying three clinical cases presented as examples of deliberate reflection and subsequently explaining the procedure on video (compared to a control group that only diagnosed clinical cases). On a post-test one to three weeks later, all participants diagnosed a new set of (test) cases while thinking aloud (to capture the residents' reasoning process; Durning et al., 2013). We hypothesized that participants in the learning-by-teaching condition would have learned and would apply deliberate reflection on the test cases, meaning they would engage in more reflective reasoning when diagnosing cases in the test phase (as indicated by the think-aloud protocols) and, therefore, would take more time to diagnose and show higher diagnostic accuracy than participants in the control condition. For additional measures on the learning process and outcome, we measured mental effort (Van Gog & Paas, 2008), an indicator of experienced cognitive load, and confidence in the given diagnosis.

## METHOD

### Participants and Design (Figure 1)

Ninety-nine residents followed our invitation and came to the first study session, and 56 of them (39 female; age *M* = 29.05, *SD* = 2.85) completed both sessions. The residents were in the first year of a three-year residency program at either the Erasmus Medical Centre in Rotterdam (*n* = 37), or the Leiden University Medical Centre (*n* = 19). The study took place during the usual educational program and participants did not receive compensation. The ethics committee of the Department of Psychology, Erasmus University Rotterdam, approved the study. Participants were randomly assigned to the learning-by-teaching condition (*n* = 28) or the control condition (*n* = 28).

Figure 1 - Illustration of the study protocol




*Note.* *The prior-experience questionnaire was only filled in by 35 of the 56 participants.

## Materials

**Prior knowledge questionnaire.** To check if there were no differences in prior knowledge between the experimental and control condition, participants filled out a prior knowledge questionnaire. Besides demographics and experience in clinical practice, participants were presented with a list of clinical symptoms and conditions (Appendix A) including those presented in the cases of this study and others (i.e. fillers) to disguise the diseases of interest. The participants were asked to indicate 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).

**Cases.** In this study we used ten written, clinical cases (Appendix B) which described complex problems as they can be encountered in general practice (example in Appendix C). Each case described a different patient with complaints, medical history, and findings from physical examination, and in some cases additional test results. The cases were prepared by experienced general practitioners. For validation, each case was solved by two different general practitioners who were blind to the intended diagnosis. If one or both general practitioners did not give the intended diagnosis, they discussed and adjusted the case until they reached agreement.



**Deliberate Reflection examples.** For the experimental condition, we used a combination of 'example-based learning' (Van Gog et al., 2019) and 'learning-by-teaching' (Fiorella & Mayer, 2013, 2014; Hoogerheide et al., 2016; Hoogerheide et al., 2014), so participants could first study examples of the procedure that they would be asked to teach. For this, three written worked-out examples were used, which illustrated how deliberate reflection was applied on a case of the learning phase (example in Appendix D). Each example showed the reflection procedure on a different case from the learning phase and analysed three plausible differential diagnoses. The deliberate reflection method aims at inducing a critical review of the initial and following diagnoses (Mamede, Schmidt, & Penaforte, 2008). The procedure requires the physician to first read a case and give an initial diagnosis. Subsequently, the physician goes back to the case and lists (1) findings that support the diagnosis, (2) findings that oppose the diagnosis, (3) findings that would have been expected if the diagnosis were true, but were absent. The physician then gives (4) an alternative diagnosis and follows the same analytical steps (1-4) for this alternative diagnosis, and for a third diagnosis. The written deliberate reflection examples left out the final step of deliberate reflection, which is the ranking of the diagnoses in order of likelihood and thereby choosing a final diagnosis, as the residents were asked to do this themselves. The three worked-out examples were prepared by experienced general practitioners.

**Mental Effort and Confidence.** To acquire additional information on the reasoning process, participants were asked to rate their mental effort when diagnosing as well as their confidence in their final diagnosis. Mental effort and confidence were each rated on different pages and on 9-point-Likert-scales ranging from 1 (very low) to 9 (very high), similar to the mental-effort rating by (Paas, 1992).

**Explanation videos.** Using a web cam recorder (www.addpipe.com), participants in the learning-by-teaching condition were instructed to record two videos in the learning phase, addressing a fictitious peer. For the first video, they were shown an empty reflection table, which had the same format as the deliberate reflection examples, but all text was removed. Participants were asked to explain what the steps of deliberate reflection are and how the procedure can help to avoid common reasoning errors. For recording the second video, they were shown one of the cases they had diagnosed earlier, together with a table containing only the steps of deliberate reflection. Participants now had to explain how the given case was diagnosed by applying the deliberate reflection procedure.

**Presentation.** The prior-knowledge questionnaire and the two study sessions were programmed in Qualtrics software (Version 05.2018). Two versions of each session presented the cases of a session in a different order to reduce the influence of item-order effects on participants' answers. During the learning phase, participants in the learning-by-teaching condition saw three cases together with the worked-out reflection examples for these cases. Participants in the control condition saw the same three cases without reflection example, and three additional cases. Each session was self-paced and participants could not move back in the program. The participants' answers and response times were saved automatically.

## Procedure

Prior to the study, participants had been told that we investigated diagnostic reasoning and the effectiveness of educational methods. Approximately two weeks before the first experimental session, the residents received a Qualtrics questionnaire per email, and were asked to fill it in prior to the session. The experimental sessions were conducted at the residents' institute and were led by different researchers all following the same instructions. At the beginning of the first session, participants were randomly distributed. Residents in the experimental group received instructions for the learning-by-teaching condition and residents in the other group for the control condition. First, all residents individually watched an instruction video on the computer which explained how an example case had been diagnosed following the instructions of their study condition. In the learning-by-teaching condition, the video therefore explained the steps of deliberate reflection and how they could

help to avoid common reasoning errors. After watching the video, all participants started with the diagnostic task.

In the control condition, participants were shown the first case. They were asked to read the case until they had decided which diagnosis is the most likely for the case and then to move on to the next page. The case disappeared from the screen and they were asked to fill in the diagnosis. On the following two pages they were asked to rate how much mental effort they invested in diagnosing the case and how much confidence they had in the diagnosis. After this, they moved on to the next case until all six cases had been diagnosed. Participants in the control condition analysed three cases more than participants in the learning-by-teaching condition. These cases had the same structure but a different content (Appendix B). As a second measure to keep the time-on-task the same for both conditions, participants in the control group then did a filler task, in which they diagnosed four unrelated internal medicine cases. These cases described patients with acute prostatitis, acute glomerulonephritis, hepatitis B and deep vein thrombosis.



Participants in the learning-by-teaching condition, were shown the first case and were asked to read it and give a diagnosis (as in the control condition). They then saw the case again along with a worked-out deliberate reflection example. Participants were asked to study this example and to rank the given diagnoses in order of likelihood. Then, they rated their mental effort and confidence, and went on to the next case until all three cases were diagnosed. When finished, participants moved on to a task wherein they recorded the two explanation videos, addressing a fictitious peer.

One to three weeks later, the test session took place (the timing difference was due to differences in the residents' class schedule). The test was the same for both conditions. Participants were asked to diagnose new cases while thinking aloud. In order to get used to the method, they did two unrelated think-aloud tasks without clinical cases. After this, they started to diagnose four new cases. Participants started the audio recorder and then saw a case. They were asked to think aloud until they had arrived at their final (most likely) diagnosis for the case. They went on to the next page and filled in this diagnosis. After this, they rated their mental effort and confidence and went on to the next case until all four cases had been diagnosed. Finally, participants received a written debriefing and were thanked for their participation.

## Data analysis

For all analyses we used a significance level of $\alpha = .05$ and did a Bonferroni correction for the number of tests, which led to $\alpha = .001$. As a measure of effect size, $\eta_p^2$ is provided for

the analyses of variances, with .01, .06, .14 corresponding to small, medium and large effects (Cohen, 1988).

**Prior knowledge.** Mean prior experience ratings were computed for the chief complaints and diagnoses of the cases. To check for initial differences between the groups, we conducted a one-way analysis of variance (ANOVA) on the mean prior experience ratings with condition (learning-by-teaching, control) as a between-subjects factor.

**Learning phase.** To check whether participants had learned the deliberate reflection procedure and whether they completed the explanation task appropriately, we analysed the first explanation video recorded under the learning-by-teaching condition, wherein residents had to explain the steps of deliberate reflection. Due to technical problems only 17 videos were recorded correctly and could be used for analysis. Two researchers independently judged whether residents named the four steps of deliberate reflection and in which order to use them. The two raters completely agreed when scoring the deliberate reflection steps and had an almost perfect interrater reliability (Landis & Koch, 1977) for scoring whether the correct order was given, *Kappa* = .87.

**Test phase.** Participants' final diagnoses were scored by two general practitioners independently as either 1 (correct core diagnosis), 0.5 (partially correct), or 0 (incorrect). The interrater reliability was excellent, *ICC* = .94 (Cicchetti, 1994), and disagreements were later resolved through discussion. Furthermore, we analysed how much time participants had spent on a case until they moved to the next page to fill in a diagnosis (time to diagnose). Participants' mean scores on the test cases were computed on diagnostic accuracy, time to diagnose, mental effort, and confidence. To analyse differences between the two conditions, we conducted a one-way ANOVA on each outcome measure.

Further, we analysed the recordings from the think-aloud task, to test whether the deliberate reflection procedure was adopted when diagnosing test cases one to three weeks later. We were missing 66 recordings (29%) due to technical errors. The remaining 158 recording from 46 participants first were transcribed. We then counted the numbers of idea units (Meyer, 1975; Schiefele & Krapp, 1996) in the think-aloud protocols. An idea unit is the smallest meaningful idea that can be identified in a fragment of text. The idea units were coded according to the deliberate reflection steps 1 -4. Thereby, a table as shown as deliberate reflection example was reconstructed from the residents' think-aloud protocols. Consequently, an idea unit that was categorised as step 1-3 could be counted more often than it was vocalised. That was the case when a resident linked one argument to multiple diagnoses. Two researchers who were blind to the condition categorised and counted the idea units without judging the correctness of the medical content. A sample of 10% of the

data was rated by both researchers with an interrater reliability ranging from fair to excellent (Cicchetti, 1994), step 1 to 4: *ICC* = .60, *ICC* = .84, *ICC* = .43, *ICC* = .68.

From these idea units we computed a measure that reflects how many key elements of deliberate reflection were used when solving a case. A crucial element is, that participants do not only consider information that supports their diagnosis at hand, but that they consider contradictory arguments and alternative diagnoses, as well, to critically reflect on their diagnosis. The aim of these steps is to help physicians to avoid a tunnel vision and confirmatory bias towards their first impression of the case, as these types of reasoning flaws have been associated with diagnostic errors (Graber et al., 2005). Therefore, we analysed the number of *contradiction units* in the participants' reasoning to measure adoption of the deliberate reflection procedure. Contradiction units were defined as the idea units that we categorised into the deliberate reflection step 2 (what speaks against), 3 (what is missing), and 4 (differential diagnoses). For the statistical analysis, the *proportion of contradiction units* was calculated relative to all idea units given by the participant (this adjusts for possible differences between cases in the total number of idea units reported). From this, we computed the participants' mean proportion of contradiction units. A one-way ANOVA was conducted on mean proportion of contradiction units with condition (learning-by-teaching, control) as a between-subjects factor.

# RESULTS

## Prior clinical experience

Table 1 shows the descriptive statistics and experience with the medical conditions and complaints. Note that only 35 of the 56 participants filled in the prior knowledge assessment. The conditions did not significantly differ on prior experience with the symptoms, $F$ (1, 33) = 2.01, $p$ = .150, $\eta_p^2$ = .06, or with the diagnoses $F$ (1, 33) = .01, $p$ = .922, $\eta_p^2$ < .01.

Table 1 - Prior experience rating of the symptoms and correct diagnoses presented in this study.

| | | All cases | |
|---|---|---|---|
| | *N* | *Mean* | *SD* |
| **Age** | | | |
| Control | 28 (17 female) | 30.21 | 3.23 |
| Learning-by-teaching | 28 (22 female) | 27.78 | 1.66 |
| Total | 56 (39 female) | 29.05 | 2.85 |
| **Prior experience with symptoms** | | | |
| Control | 20 | 3.05 | .43 |
| Learning-by-teaching | 15 | 3.25 | .39 |
| Total | 35 | 3.14 | .42 |
| **Prior experience with diagnoses** | | | |
| Control | 20 | 2.57 | .59 |
| Learning-by-teaching | 15 | 2.55 | .51 |
| Total | 35 | 2.56 | .55 |

*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).

## Learning phase

Out of the 17 explanation videos we analysed, 11 residents described the procedure perfectly. Four residents described all steps but did not state clearly that you should first analyse one diagnosis and only after that think of the next diagnosis. This might be important for deliberate reflection to be effective (Mamede & Schmidt, 2014). Two residents did not state clearly that when falsifying one's diagnosis you should include symptoms that you would have expected if the diagnosis was true but were absent in the case.

## Test phase

Table 2 shows the mean and standard deviation of diagnostic accuracy, mental effort, confidence, time to diagnose, and proportion of contradiction units that were measured during the post-test. One-way ANOVAs showed no main effects of condition on diagnostic accuracy, $F$ (1, 54) = 1.28, $p$ = .263, $\eta_p^2$ = .02, on time to diagnose, $F$ (1, 54) = .28, $p$ = .598, $\eta_p^2$ < .00, on mental effort ratings, $F$ (1, 54) = .37, $p$ = .544, $\eta_p^2$ = .01, on confidence ratings, $F$ (1, 54) = .14, $p$ = .710, $\eta_p^2$ < .01, or on the proportion of contradiction units, $F$ (1, 43) = .37, $p$ = .544, $\eta_p^2$ = .01. Note that only 45 residents were included in the latter analysis because for 11 residents the think-aloud task was not recorded correctly.

Table 2 - All outcome measures collected during the post-test.

| | | All cases | |
|---|---|---|---|
| | *N* | *Mean* | *SD* |
| **Diagnostic Accuracy** | | | |
| Control | 28 | .51 | .30 |
| Learning-by-teaching | 28 | .59 | .23 |
| Total | 56 | .55 | .27 |
| **Time to Diagnose** | | | |
| Control | 28 | 271.00 | 86.41 |
| Learning-by-teaching | 28 | 260.64 | 56.68 |
| Total | 56 | 265.82 | 72.59 |
| **Mental Effort** | | | |
| Control | 28 | 5.40 | 1.37 |
| Learning-by-teaching | 28 | 5.59 | 1.01 |
| Total | 56 | 5.49 | 1.20 |
| **Confidence** | | | |
| Control | 28 | 5.62 | 1.00 |
| Learning-by-teaching | 28 | 5.72 | 1.14 |
| Total | 56 | 5.67 | 1.06 |
| **Proportion of Contradiction Units** | | | |
| Control | 22 | .29 | .10 |
| Learning-by-teaching | 24 | .31 | .09 |
| Total | 46 | .30 | .10 |

*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).

## DISCUSSION

Although prior studies have shown that deliberate reflection improves diagnosis (Mamede, Schmidt, et al., 2010; Mamede, Schmidt, & Penaforte, 2008; Mamede, Van Gog, et al., 2010; Schmidt et al., 2014; Schmidt et al., 2017), it is as yet unclear whether the deliberate reflection procedure can be learned and autonomously applied on future cases (without prompting physicians to do so). We therefore investigated whether general practice residents would learn the deliberate reflection procedure by studying examples and explaining the procedure on video (compared to a control group that only diagnosed cases) and apply it when solving novel cases one to three weeks later. There were no differences between the learning-by-teaching condition and the control condition in the proportion of contradictory idea units reported while diagnosing the case, time needed to diagnose, and diagnostic accuracy. Practicing with deliberate reflection also did not influence participants' confidence in their diagnosis or mental effort needed to solve future cases. Against our expectations, these findings suggest that the two conditions did not differ in the extent to which they incorporated elements of the deliberate reflection procedure in their reasoning process.

One possible explanation is, that all the residents already naturally engaged in reflective reasoning. The cases in this study were designed to be difficult and to be more complex than in clinical practice, because complexity is known to trigger reflective reasoning (Mamede, Schmidt, et al., 2010; Mamede, Schmidt, Rikers, et al., 2008; Mamede et al., 2007). Moreover, the request to diagnose the cases while thinking aloud probably also induced a more thorough consideration of the case than what they would naturally do. Perhaps these cases stimulated reflection in all the residents. If the residents in the control condition reasoned similarly to those in the learning-by-teaching condition, who had learned which reasoning steps help them to prevent errors, this means that the residents could already engage in some sort of reflective reasoning. Therefore, deliberate reflection might not further improve the residents' diagnostic reasoning. This explanation is supported by comments from the residents' teachers who said that they always expect their trainees to generate multiple differential diagnoses for a case. Thus, it is possible that their education already implies the steps of deliberate reflection to some degree and that residents in this phase of postgraduate training are able to reflect and therefore engage in reflective reasoning when solving cases that are not straightforward.

An alternative explanation is that residents in the learning-by-teaching condition did learn the deliberate reflection procedure but did not apply it during the test phase. The videos of the first explanation task suggest that the residents had learned the deliberate reflection procedure. However, in order to adopt it as a diagnostic strategy for themselves, perhaps they would need more practice with the procedure (i.e., automatize it), with a shorter time

interval between the sessions. Future studies could test whether a learning phase with multiple sessions would be effective for residents in adopting deliberate reflection. In contrast to prior studies (Fiorella & Mayer, 2013; Hoogerheide, Renkl, et al., 2019) the participants in this study did not have fixed times to study or explain the learned material. We do not know whether a fixed study period would have helped participants to make better use of their study opportunity. It may also be that the residents did not feel the need to engage in reflection. As we explained above, the cases were prepared to be difficult, because higher difficulty levels tend to trigger reflection (Mamede, Schmidt, Rikers, et al., 2008; Mamede et al., 2007). The residents' diagnostic accuracy showed to be at an intermediate level, at which deliberate reflection has been shown to be beneficial (Costa Filho et al., 2019). However, it is also known that physicians' perception of how difficult a case is far from an objective, accurate judgement (Meyer et al., 2013). Perhaps the residents in our study did not perceive the cases as demanding enough to require further thinking.



Another explanation is, that though residents in the learning-by-teaching condition did learn the deliberate reflection procedure, this does not mean that they have learned to adopt the procedure as a general reasoning process for addressing future problems. While cognitive interventions can improve diagnostic accuracy when physicians are explicitly instructed to use them (Lambe et al., 2016; Prakash et al., 2019) it has been questioned whether generalizable cognitive skills that could be applied to new problems, can be taught (Eva et al., 1998; Monteiro et al., 2020; Norman, 1988). Content specific interventions that increase or reorganize medical knowledge may be more effective to improve diagnostic accuracy (Norman et al., 2017; Schmidt & Mamede, 2015). Hoogerheide, Renkl, et al. (2019) may have found transfer of the learned problem-solving procedure to novel problems because their learning problems and test problems were more similar in content than the different cases in the present study were. A limitation of the study is the substantial drop-out from the first to the second session, which reduced our sample size. The missing think-aloud data, that could not be analysed, further reduced our sample size, which may have caused the study power to be insufficient to find an existing effect. Besides that, we do not know whether the think-aloud task in the test phase affected the residents' reasoning and fostered reflective reasoning of all residents. Being required to think aloud while reasoning naturally leads to considering case findings more extensively, eventually 'removing' physicians from an intuitive reasoning mode. Furthermore, we do not know whether four cases in the test phase were enough to find a possible difference between the conditions. Another limitation is that we have no objective standard of what can be considered much or little reflection. As both conditions performed the same, we cannot say whether this is because both engaged in much or little reflective reasoning. Future studies should include a reflection template to which the participants' reasoning can be compared. Furthermore, qualitative studies could give more insight into the reasoning process.

Given that our residents might perhaps already have had too much experience with reflection, it would be interesting for future research to test whether learning-by-teaching, which seems to be particularly effective for students with little prior knowledge (Hoogerheide, Renkl, et al., 2019), would be effective to teach deliberate reflection to medical students. Ibiapina et al. (2014) conducted a study among students in which they focused on effects of deliberate reflection on learning about the content knowledge of the cases. In contrast to our results, they found that practicing with deliberate reflection increased diagnostic accuracy on cases diagnosed one week later. In that study the future test cases were similar to the practice cases, whereas in our study we also included unrelated test cases. Therefore, it can be that the benefit of deliberate reflection on improving future diagnostic accuracy is only due to learning about the specific content of the cases rather than the reflective procedure, and does not transfer to cases with unrelated diseases. However, Ibiapina et al. did not test the effect on unrelated cases and we do not know whether it also had an effect on the students' reasoning process. Future studies should conduct the present study with students, to see whether practicing with deliberate reflection is effective in teaching reflective reasoning if participants are less experienced than residents are.

To sum up, the results of the present study showed that for residents in the general practice training, practicing with deliberate reflection by explaining it on video did not increase reflective reasoning on future cases. It could be that the residents did not yet adopt the procedure and that more practice is needed, or that the residents did not feel the need to apply the procedure in the test phase. Another explanation is that the control condition also engaged in reflective reasoning during the test phase, and that the added benefit of deliberate reflection is too small to find an effect.

## REFERENCES


Atkinson, R. K., Derry, S. J., Renkl, A., & Wortham, D. (2000). Learning from Examples: Instructional Principles from the Worked Examples Research. *Review of Educational Research*, *70*(2), 181-214.

Brown, A. L., & Kane, M. J. (1988). Preschool children can learn to transfer: Learning to learn and learning from example. *Cognitive Psychology*, *20*(4), 493-523. https://doi.org/https://doi.org/10.1016/0010-0285(88)90014-X

Cicchetti, D. V. (1994). Guidelines, Criteria, and Rules of Thumb for Evaluating Normed and Standardized Assessment Instruments in Psychology. *Psychological Assessment*, *6*(4), 284-290.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2 ed.). Lawrence Erlbaum Associates.

Costa Filho, G. B., Moura, A. S., Brandão, P. R., Schmidt, H. G., & Mamede, S. (2019). Effects of deliberate reflection on diagnostic accuracy, confidence and diagnostic calibration in dermatology. *Perspectives on Medical Education*, *8*(4), 230-236. https://doi.org/10.1007/s40037-019-0522-5

Dewey, J. (1910). *How we think*. D.C. Heath & Co.

Durning, S. J., Artino, A. R., Beckman, T. J., Graner, J., der Vleuten, C. v., Holmboe, E., & Schuwirth, L. (2013). Does the think-aloud protocol reflect thinking? Exploring functional neuroimaging differences with thinking (answering multiple choice questions) versus thinking aloud. *Medical Teacher*, *35*(9), 720-726. https://doi.org/10.3109/0142159x.2013.801938

Eva, K. W., Neville, A. J., & Norman, G. R. (1998). Exploring the etiology of content specificity: factors influencing analogic transfer and problem solving. *Academic Medicine*, *73*(10 Suppl), S1-5.

Fiorella, L., & Mayer, R. E. (2013). The relative benefits of learning by teaching and teaching expectancy. *Contemporary Educational Psychology*, *38*(4), 281-288. https://doi.org/10.1016/j.cedpsych.2013.06.001

Fiorella, L., & Mayer, R. E. (2014). Role of expectations and explanations in learning by teaching. *Contemporary Educational Psychology*, *39*(2), 75-85. https://doi.org/10.1016/j.cedpsych.2014.01.001

Graber, M. L., Franklin, N., & Gordon, R. (2005). Diagnostic Error in Internal Medicine. *Archives of Internal Medicine*, *165*(13), 1493-1499. https://doi.org/10.1001/archinte.165.13.1493

Hoogerheide, V., Deijkers, L., Loyens, S. M., & Heijltjes, A. (2016). Gaining from explaining: Learning improves from explaining to fictitious others on video, not from writing to them. *Contemporary Educational Psychology*, *44*, 95-106. https://doi.org/http://dx.doi.org/10.1016/j.cedpsych.2016.02.005

Hoogerheide, V., Loyens, S. M. M., & Van Gog, T. (2014). Effects of creating video-based modeling examples on learning and transfer. *Learning and Instruction*, *33*, 108-119. https://doi.org/10.1016/j.learninstruc.2014.04.005

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. *Journal of Educational Psychology*, *111*(1), 45-56.

Hoogerheide, V., Visee, J., Lachner, A., & Van Gog, T. (2019). Generating an instructional video as homework activity is both effective and enjoyable. *Learning and Instruction*, *64*, 101226. https://doi.org/https://doi.org/10.1016/j.learninstruc.2019.101226

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. *Medical Education*, *48*, 796-805. https://doi.org/10.1111/medu.12435

Koh, A. W. L., Lee, S. C., & Lim, S. W. H. (2018). The learning benefits of teaching: A retrieval practice hypothesis. *Applied Cognitive Psychology*, *32*(3), 401-410. https://doi.org/https://doi.org/10.1002/acp.3410

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 Professions Education*, *6*(4), 506-515. https://doi.org/https://doi.org/10.1016/j.hpe.2020.07.004

Lambe, K. A., Reilly, G., Kelly, B. D., & Curristan, S. (2016). Dual-process cognitive interventions to enhance diagnostic reasoning: a systematic review. *BMJ Quality & Safety*, *25*(10), 808. https://doi.org/10.1136/bmjqs-2015-004417

Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, *33*(1), 159-174.

Mamede, S., Schmidt, H., Rikers, R., Custers, E., Splinter, T., & Saase, J. (2010). Conscious thought beats deliberation without attention in diagnostic decision-making: At least when you are an expert. *Psychological Research*, *74*(6), 586-592.

Mamede, S., & Schmidt, H. G. (2014). Reflection in Diagnostic Reasoning: What Really Matters? *Academic Medicine*, *89*(7), 959-960. https://doi.org/10.1097/acm.0000000000000306

Mamede, S., Schmidt, H. G., & Penaforte, J. C. (2008). Effects of reflective practice on the accuracy of medical diagnoses. *Medical Education*, *42*(5), 468-475. https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1365-2923.2008.03030.x

Mamede, S., Schmidt, H. G., Rikers, R. M., Penaforte, J. C., & Coelho-Filho, J. M. (2008). Influence of perceived difficulty of cases on physicians' diagnostic reasoning. *Academic Medicine*, *83*(12), 1210-1216. https://doi.org/10.1097/ACM.0b013e31818c71d7

Mamede, S., Schmidt, H. G., Rikers, R. M. J. P., Penaforte, J. C., & Coelho-Filho, J. M. (2007). Breaking down automaticity: Case ambiguity and the shift to reflective approaches in clinical reasoning. *Medical Education*, *41*(12), 1185-1192. https://doi.org/10.1111/j.1365-2923.2007.02921.x

Mamede, S., Van Gog, T., Van den Berge, K., Rikers, R. M., Van Saase, J. L., Van Guldener, C., & Schmidt, H. G. (2010). Effect of Availability Bias and Reflective Reasoning on Diagnostic Accuracy Among InternalMedicine Residents. *JAMA*, *304*(11), 1198-1203. https://doi.

org/10.1001/jama.2010.1276

Mann, K., Gordon, J., & MacLeod, A. (2009). Reflection and reflective practice in health professions education: a systematic review. *Advances in Health Science Education*, *14*(4), 595-621.

Meyer, A. N. D., Payne, V. L., Meeks, D. W., Rao, R., & Singh, H. (2013). Physicians' diagnostic accuracy, confidence, and resource requests: a vignette study. *JAMA Internal Medicine*, *173*(21), 1952-1958. https://doi.org/10.1001/jamainternmed.2013.10081

Meyer, B. J. F. (1975). *The organization of prose and its effects on memory*. Amsterdam: North-Holland Publishing Co.

New York: American Elsevier Publishing Co.

Monteiro, S., Sherbino, J., Sibbald, M., & Norman, G. (2020). Critical thinking, biases and dual processing: The enduring myth of generalisable skills. *Medical Education*, *54*(1), 66-73. https://doi.org/https://doi.org/10.1111/medu.13872

Ng, S. L., Kinsella, E. A., Friesen, F., & Hodges, B. (2015). Reclaiming a theoretical orientation to reflection in medical education research: a critical narrative review. *Medical Education*, *49*(5), 461-475.

Norman, G. R. (1988). Problem-solving skills, solving problems and problem-based learning. *Medical Education*, *22*(4), 279-286. https://doi.org/https://doi.org/10.1111/j.1365-2923.1988.tb00754.x

Norman, G. R., Monteiro, S., Sherbino, J., Ilgen, J. S., Schmidt, H. G., & Mamede, S. (2017). The Causes of Errors in Clinical Reasoning: Cognitive Biases, Knowledge Deficits, and Dual Process Thinking. *Academic Medicine*, *92*(1), 23-30. https://doi.org/10.1097/acm.0000000000001421

Paas, F. (1992). Training Strategies for Attaining Transfer of Problem-Solving Skill in Statistics: A Cognitive-Load Approach. *Journal of Educational Psychology*, *84*(4), 429-434. https://doi.org/10.1037/0022-0663.84.4.429

Prakash, S., Sladek, R. M., & Schuwirth, L. (2019). Interventions to improve diagnostic decision making: A systematic review and meta-analysis on reflective strategies. *Medical Teacher*, *41*(5), 517-524. https://doi.org/10.1080/0142159x.2018.1497786

Roediger, H. L., & Karpicke, J. D. (2006). The Power of Testing Memory: Basic Research and Implications for Educational Practice. *Perspectives on Psychological Science*, *1*(3), 181-210.

Rummel, N., & Spada, H. (2005). Learning to Collaborate: An Instructional Approach to Promoting Collaborative Problem Solving in Computer-Mediated Settings. *Journal of the Learning Sciences*, *14*(2), 201-241.

Schiefele, U., & Krapp, A. (1996). Topic interest and free recall of expository text [doi:10.1016/S1041-6080(96)90030-8]. *Learning and Individual Differences*, *8*(2), 141-160. https://doi.org/10.1016/s1041-6080(96)90030-8

Schmidt, H. G., & Mamede, S. I. (2015). How to improve the teaching of clinical reasoning: a narrative review and a proposal. *Medical Education*, *49*(10), 961-973. https://doi.org/10.1111/medu.12775

Schmidt, H. G., Mamede, S. I., Van Den Berge, K., Van Gog, T., Van Saase, J. L. C. M., & Rikers,

R. M. J. P. (2014). Exposure to media information about a disease can cause doctors to misdiagnose similar-looking clinical cases. *Academic Medicine*, *89*(2).

Schmidt, H. G., Van Gog, T., Schuit, S. C. E., Van Den Berge, K., Van Daele, P. L. A., Bueving, H., Van der Zee, T., Van Den Broek, W. W., Van Saase, J. L. C. M., & Mamede, S. I. (2017). Do patients' disruptive behaviours influence the accuracy of a doctor's diagnosis? A randomised experiment. *BMJ Quality & Safety*, *26*(1), v19-23. https://doi.org/10.1136/bmjqs-2015-004109

Van Gog, T., & Paas, F. (2008). Instructional Efficiency: Revisiting the Original Construct in Educational Research [doi: 10.1080/00461520701756248]. *Educational Psychologist*, *43*(1), 16-26. https://doi.org/10.1080/00461520701756248

Van Gog, T., & Rummel, N. (2010). Example-Based Learning: Integrating Cognitive and Social-Cognitive Research Perspectives. *Educational Psychology Review*, *22*(2), 155-174. https://doi.org/10.1007/s10648-010-9134-7

Van Gog, T., Rummel, N., & Renkl, A. (2019). Learning How to Solve Problems by Studying Examples. In J. Dunlosky & K. A. Rawson (Eds.), *The Cambridge Handbook of Cognition and Education* (pp. 183-208). Cambridge University Press. https://doi.org/10.1017/9781108235631.009

# APPENDIX A

*Diagnoses and symptoms shown during the prior knowledge assessment.*

| Correct diagnoses and chief complaints that appear in the cases | Other diagnoses and symptoms (filler) |
|---|---|
| Diarrhoea | Constipation |
| Irritable bowel syndrome (IBS) | Abdominal pain |
| Inflammatory bowel disease (IBD) | Gastroenteritis |
| Chronic pancreatitis | Infection of the upper respiratory tract |
| Shortness of breath | Chronic Obstructive Pulmonary Disease |
| Infection of the lower respiratory tract | Anaemia |
| Pulmonary embolism | Pain on the chest |
| Rash in the face | Depression |
| Rosacea | Eating disorder |
| Tremor in hand | Vaginal complaints |
| Multiple sclerosis | Vaginal fungal infection |
| Facial paralysis | Sexually Transmitted Diseases |
| Idiopathic peripheral facial paralysis (IPAV) | Erectile dysfunction |
| | Headache |
| | Dizziness |
| | Thumb base instability |
| | Shingles |
| | Acne |
| | Pain in thumb |
| | Weak muscles |
| | Hyperthyroidism |
| | Cerebrovascular Accident (CVA) |
| | Lyme disease |
| | Diverticulitis |
| | Asthma |
| | Heart failure |
| | Palpitations |
| | Anxiety / panic disorder |
| | Vaginal discharge |
| | Amenorrhea |
| | Bacterial vaginosis |
| | Benign Paroxysmal Position Vertigo (BPPD) |
| | Pregnancy |
| | Skin rash |
| | Scarlet fever |
| | Lower back pain |

| | |
|---|---|
| | Pain in legs |
| | Spondylodiscitis |
| | Spinal canal stenosis / neurogenic claudication |
| | Turn dizziness |

## APPENDIX B

*Overview of the medical conditions of the cases used during the different session. Chief complaints are given in parenthesis. Three cases in the learning phase have only been shown to the control condition and are market with an asterisk.*

| Learning session | Post-test session |
|---|---|
| Chronic pancreatitis (diarrhoea) | Chronic pancreatitis (diarrhoea) |
| Inflammatory bowel disease (diarrhoea) | Inflammatory bowel disease (diarrhoea) |
| Irritable bowel syndrome (diarrhoea) | Infection of the lower respiratory tract (shortness of breath) |
| Bell's palsy* (facial paralysis) | Pulmonary embolism (shortness of breath) |
| Rosacea* (Rash in the face) | |
| Multiple sclerosis* (tremor in hand) | |

*Note. Each case was different, i.e. described a different patient.*

## APPENDIX C

*Translation of a case of irritable bowel syndrome, with as chief complaint diarrhoea. This case has been shown during the learning session.*

You haven't seen Mrs. Alkema (27 years old) in a while. Since the last time you spoke with her, she completed her law studies and is working long days at a law firm since. She enjoys her work. She likes to really delve into a case and get the most out of it. Recently she moved in a new house in a village together with her girlfriend, and every morning she gets into the car to join the traffic jam towards the city. But this morning she has an appointment at your practice and she seems slightly irritated that your schedule is 10 minutes delayed.



She quickly explains why she came: severe abdominal pain. But really severe; so bad that at these moments she just has to go to the bathroom. Even when she is in a meeting. Once she even had to leave the courtroom, for which she feels embarrassed. However, it always brings her some relief for a moment. At these times, she has somewhat thin stool. Occasionally there is some blood on it. You ask where the abdominal pain is located; that is clear, around the navel.

"This can't go on; can't you just refer me to a specialist?" When you ask her, it appears that she has been suffering from this for more than half a year. To keep going, she has used Naproxen 250mg twice a day for the last 2 weeks. Her menstruation is not as much as it used to be, sometimes it skips. To lose weight, she has been following a diet for the last 2 months, which you have not heard of before. It is working; she has lost some weight. At the end of the consultation, you remember that her mother also often had abdominal pain.

**Physical examination:**

During the physical examination, you find pressure pain near the descending colon, but no further abnormalities. The checks are good.

**Additional tests:**

You decide to order laboratory tests, in the hope that the results will be reassuring. You do not have the results yet.

## APPENDIX D

*Translation of the worked-out deliberate reflection example shown to participants in the learning-by-teaching condition during the learning session.*

| Diagnosis | Which findings from the case speak **for** this diagnosis? | Which findings from the case speak **against** this diagnosis? | Which further findings would you expect if your diagnosis was correct, which are **missing** for this patient? | Finally: Order of likelihood (1 = most likely) |
|---|---|---|---|---|
| Inflammatory bowel disease (IBD) | - Diarrhoea with intervals<br>- Rectal blood loss | - Complaints always temporary<br>- No involuntary weight loss or other additional complaints | - Alarm signals (weight loss, tired)<br>- Perianal abnormality<br>- Fever<br>- Abdominal resistance<br>- Family members with IBD | |
| Irritable bowel syndrome (IBS) | - Duration and frequency of abdominal pain and diarrhoea<br>- Complaints reduce after defecation<br>- Complaints always come with diarrhoea<br>- Age <50 years<br>- Hard working (stress)<br>- Abdominal pain around navel<br>- Pressure pain colon descendens<br>- Mother too had stomach ache | - Rectal blood loss (but probably hemorrhoid) | - Mucus with stool<br>- Flatulence<br>- Depending on food / stress | |
| Diverticulitis (complicated/ uncom-plicated) | - Pressure pain in left lower abdomen<br>- Recurrent course<br>- No alarm symptoms<br>- Rectal blood loss | | - More and sharper pain<br>- More pressure and release pain<br>- Fever<br>- Abdominal resistance | |

# CHAPTER 5

# Improving Medical Residents' Self-Assessment of Their Diagnostic Accuracy: Does Feedback Help?

**This chapter has been published as:**

Kuhn, J., Van den Berg, P., Mamede, S., Zwaan, L., Bindels, P., & Van Gog, T. (2022). Improving medical residents' self-assessment of their diagnostic accuracy: Does feedback help? *Advances in Health Sciences Education, 27*(1), 189–200. https://doi.org/10.1007/s10459-021-10080-9

## ABSTRACT

When physicians do not estimate their diagnostic accuracy correctly, i.e. show inaccurate *diagnostic calibration*, diagnostic errors or overtesting can occur. A previous study showed that physicians' diagnostic calibration for easy cases improved, after they received feedback on their previous diagnoses. We investigated whether diagnostic calibration would also improve from this feedback when cases were more difficult. Sixty-nine general-practice residents were randomly assigned to one of two conditions. In the feedback condition, they diagnosed a case, rated their confidence in their diagnosis, their invested mental effort, and case complexity, and then were shown the correct diagnosis (feedback). This was repeated for 12 cases. Participants in the control condition did the same without receiving feedback. We analysed calibration in terms of (1) absolute accuracy (absolute difference between diagnostic accuracy and confidence), and (2) bias (confidence minus diagnostic calibration). There was no difference between the conditions in the measurements of calibration (absolute accuracy, $p = .204$; bias, $p = .176$). Post-hoc analyses showed that on correctly diagnosed cases (on which participants are either accurate or underconfident), calibration in the feedback condition was less accurate than in the control condition, $p = .013$. This study shows that feedback on diagnostic performance did not improve physicians' calibration for more difficult cases. One explanation could be that participants were confronted with their mistakes and thereafter lowered their confidence ratings even if cases were diagnosed correctly. This shows how difficult it is to improve diagnostic calibration, which is important to prevent diagnostic errors or maltreatment.

## INTRODUCTION

Physicians do not always estimate their diagnostic performance correctly (Costa Filho et al., 2019; Davis et al., 2006; Friedman et al., 2005; Meyer et al., 2013). This inaccurate *diagnostic calibration* (Meyer et al., 2013), the mismatch between diagnostic accuracy and confidence in that diagnosis, can have harmful effects for the patient. Although diagnostic errors can have many causes, including system-related causes, cognitive errors play a substantial role. For example, a review of diagnostic errors in internal medicine (Graber et al., 2005) has estimated that cognitive factors play a role in around 74 % of these cases. On the one hand, being too confident in one's diagnosis might lead to *premature closure* (which is often found to occur in cases of cognitive errors; Berner & Graber, 2008; Graber et al., 2005), where physicians stop considering alternative diagnoses too early. Overconfidence has also been linked to decreased requests for diagnostic tests (Meyer et al., 2013). Being underconfident (i.e., unnecessarily uncertain) in a correct diagnosis, on the other hand, could lead to unnecessary further testing and lengthen the diagnostic process. Furthermore, the ability to correctly self-asses one's performance can help to identify potential learning needs (see self-regulated learning; Zimmerman, 2008). Improving diagnostic calibration, therefore, could not only help to prevent diagnostic errors but could also aid physicians' lifelong learning and allow them to become better performers (Eva & Regehr, 2005; Meyer & Singh, 2019; Zwaan & Hautz, 2019).



Studies from cognitive psychology have shown, that calibration of self-assessments made after performance (Hacker et al., 2008) can be improved by providing students with feedback on their previous performance (Labuhn et al., 2010; Lipko et al., 2009; Nederhand et al., 2019; Rawson & Dunlosky, 2007). The same may be true for improving calibration in a medical context: A study by Nederhand et al. (2018) showed that feedback on previous diagnostic performance improved future diagnostic calibration for medical experts as well as for medical students. In that study, participants diagnosed three cases and rated their confidence, after which some of them got feedback for the case in the form of performance standards, i.e. the correct diagnosis, and others did not get feedback. Subsequently, all physicians took the same test where they diagnosed three new, unrelated cases and rated their confidence. It was found that physicians who had previously received feedback on their diagnostic performance showed better diagnostic calibration on the test cases. However, in this study, they used relatively easy cases (resulting in high diagnostic accuracy) and it has been found that physicians' calibration is less accurate for difficult cases than for easy cases (Meyer et al., 2013).

Therefore, improving calibration on difficult cases would be even more important in order to prevent diagnostic errors. In clinical practice, physicians do sometimes get feedback in the

form of clinician report cards that show some of their performance measures in comparison to colleagues, e.g. mortality after surgery (Shahian et al., 2001). These cards have been found to help physicians improve some medical outcomes (see for example Kahi et al., 2013), but they do not yet exist for improving the diagnostic process. If feedback on diagnostic accuracy would improve diagnostic calibration, it would be valuable to use diagnostic report cards as well. Furthermore, feedback could possibly help as an educational tool for physicians in training to identify their learning needs and learn to estimate their performance better. Less over- and underconfidence in physicians in training, could potentially prevent future errors in clinical practice. In the current study, we aimed to investigate whether feedback (providing the correct diagnosis) can help to improve diagnostic calibration for residents in general practice (GP), i.e. physicians in training to become specialist, when cases are more difficult. Thus, we wanted to test whether the findings by Nederhand et al. (2018) would also show with different cases and participants in a slightly different design. Residents were asked to diagnose a case, rate their confidence in the diagnosis, and then either got the correct diagnosis for the case or moved on to the next case without feedback. We expected that GP residents who got feedback would show more accurate diagnostic calibration than residents who did not get the feedback. Additionally, we measured perceived mental effort when diagnosing the cases as well as perceived case complexity to check that the cases were not (perceived as) too easy.

# METHOD

## Participants

Ninety-seven residents in their first year of the three-year general practice training at the department of general practice at the Erasmus Medical Centre, Rotterdam, were invited to participate in this study. Sixty-nine of them accepted the invitation and completed the session (54 female; age $M$ = 29.29, $SD$ = 2.51). The study took place during the usual educational program and participants did not receive compensation.

## Material

Twelve written cases were used in this study, describing different patients with different medical conditions (Appendix A). The cases were prepared and validated by experienced general practitioners, and used in previous studies (Kuhn et al., 2020). The study was programmed in Qualtrics software (version 05.2019). For each condition, we made six versions of the program, which presented the cases in different orders. Participants moved through the program self-paced and could only move forwards. Qualtrics automatically recorded the participants' answers.



## Design and Procedure

The study was conducted in one session in computer rooms at the Erasmus Medical Centre. First, participants were asked to read the information letter on their desk and give written informed consent. Another sheet of paper provided a URL that led to one of the 12 Qualtrics programs. These papers were distributed throughout the room, so that participants were randomly assigned to either the feedback condition ($n$ = 34) or the no-feedback (i.e. control) condition ($n$ = 35). In the program, they received all instructions required for their condition together with an example case to get acquainted with the procedure. After that, they started diagnosing the first of the twelve cases.

**Feedback condition.** Participants were shown a case and were asked to read it until they had arrived at one most likely diagnosis. They moved on to the next page where they had to fill in their diagnosis. On the next three pages, they were asked to rate their confidence in their diagnosis, their mental effort invested in solving the case, and the complexity of the case. Those 3 measures were rated on 9-point-Likert scales ranging from 1 (very, very little) to 9 (very, very much). Mental effort and complexity were both used as indicators of how complex the cases were for participants. On the next page, participants were shown the correct diagnosis for the case together with the diagnosis they themselves had given and

were asked to compare both diagnoses. When they confirmed that they had compared them, they were able to move on to the next case until all twelve cases had been diagnosed.

After completing the 12 cases, participants were asked about their demographics and prior experience. They were shown a list of the diseases and chief symptoms/complaints that were used in this study, and were asked to rate their prior experience on a 5 point-Likert scale ranging from 1 (I have never seen a patient with this disease, symptom or complaint) to 5 (I have already seen many patients with this disease, symptom or complaint). Finally, participants were given a written debriefing and thanked for their time and effort.

**Control condition.** Participants in the no-feedback control condition followed the same procedure as those in the feedback condition, except they did not receive the information on the correct diagnosis for the case and the request to compare it with their own diagnosis.

## Analysis

The data were analysed using IBM SPSS Statistics 25 for Windows. For all analyses we used a significance level of $\alpha = .05$. As a measure of effect size, $\eta_p^2$ is provided for the analyses of variances, with .01, .06, .14 corresponding to small, medium and large effects (Cohen, 1988).

**Prior experience.** To analyse potential differences in prior experience between the conditions, we computed the mean prior experience ratings for the symptoms and diagnoses used in this study. On both variables, we conducted an ANOVA with condition (feedback/no feedback) as a between-subjects factor.

**Calibration.** Experienced general practitioners independently rated the diagnostic accuracy of the given diagnoses while blinded for the experimental condition, assigning either 1 (correct), .5 (partly correct), or 0 (incorrect) points. Each diagnosis was rated by two general practitioners with an 'excellent' interrater reliability, *ICC* = .96 (Cicchetti, 1994). Afterwards, they would come together and discuss the diagnoses where they had not given the same score until they reached agreement, so that each diagnosis had only one score. To calculate diagnostic calibration, we transformed the confidence ratings to match the scale of the diagnostic accuracy scores (cf. Nederhand et al., 2018): Confidence scores 1 - 3 were recoded into 0, 4 - 6 into .5, and 7 - 9 into 1. This adjustment also took into account that participants are usually reluctant to use extreme response on a Likert scale (i.e. central tendency bias).

We then computed calibration in terms of *absolute accuracy* and *bias* measures by subtracting the diagnostic accuracy scores from the transformed confidence ratings (Griffin et al., 2019). Absolute accuracy is the absolute (i.e., unsigned) difference between the two and ranges from 0 (perfect calibration) to 1 (fully inaccurate). Bias is the signed difference between the two and ranges from +1 (complete overestimation) to -1 (complete underestimation) with 0 again meaning perfect calibration. Per participant, we calculated the mean absolute accuracy and bias scores across all 12 cases. On both outcome measures, we performed an ANOVA with condition as a between-subject variable. Also, we performed a t-test on mean bias to see if it significantly differed from 0 (i.e., as zero means correct calibration, this analysis will tell whether there was significant underestimation or overestimation).

**Post hoc exploratory analyses.** In an exploratory analysis we took a closer look at calibration in relation to diagnostic accuracy. For each participant, we computed the mean bias on cases diagnosed incorrectly (diagnostic accuracy = 0; cases $n$ = 473) and on cases diagnosed correctly (diagnostic accuracy = 1; cases $n$ = 341). This may give more insight into differences in overconfidence and underconfidence between the conditions than averaging over the 12 cases. That is, on incorrectly diagnosed cases, participants will either be accurate or overconfident, whereas on correctly diagnosed cases they will either be accurate or underconfident (so by computing the mean bias across the 12 cases, overconfidence and underconfidence might cancel each other out). Note that these means were based on a different number of cases for each participant, depending on the individual performance. Partly correct cases (diagnostic accuracy = .5; cases $n$ = 14) were left out of this analysis. We performed separate ANOVAs for correct and incorrect cases, with condition as a between-subjects factor.

# RESULTS

## Prior-experience ratings

Appendix B shows the demographics and mean prior experience ratings. The analyses showed no differences between the conditions on mean prior-experience ratings for the diagnoses, $F(1, 67) = 0.12$, $p = .727$ $\eta_p^2 < .01$, and the symptoms, $F(1, 67) = .05$, $p = .831$, $\eta_p^2 < .01$, that were used in the cases of this study.

## Descriptive Statistics

Table 1 shows the means for all outcome measures (diagnostic accuracy, confidence, complexity, absolute accuracy, bias). Mean diagnostic accuracy ($M = .42$), and mean confidence ($M = 5.63$), mental effort ($M = 5.07$) and complexity ($M = 5.52$) ratings, were at an intermediate level and showed no ceiling- or floor effects.

Table 1 - Mean and standard deviation for all outcome measures (diagnostic accuracy, confidence in the diagnosis, mental effort, case complexity, and as measures of calibration: absolute accuracy and bias).

| | No-feedback condition (n = 35) | | Feedback condition (n = 34) | | Total (n = 69) | |
|---|---|---|---|---|---|---|
| | Mean | SD | Mean | SD | Mean | SD |
| Diagnostic accuracy | .42 | .14 | .43 | .12 | .42 | .13 |
| Confidence rating | 5.82 | .80 | 5.43 | .79 | 5.63 | .82 |
| Mental effort rating | 5.02 | 1.04 | 5.12 | .90 | 5.07 | .97 |
| Complexity rating | 5.64 | .82 | 5.40 | .89 | 5.52 | .86 |
| Absolute accuracy | .42 | .12 | .46 | .09 | .44 | .11 |
| Bias | .22 | .21 | .15 | .20 | .18 | .21 |

*Note.* Diagnostic accuracy was scored as either 0 (incorrect), 0.5 (partially correct) or 1 (correct). Confidence and complexity were rated on a 9-point Likert-scale ranging from 1 (very, very low) to 9 (very, very high). Absolute accuracy ranges from 0 to 1. Bias ranges from -1 to +1.

## Calibration Accuracy and Bias

The analysis of calibration on all 12 cases[1], showed no effect of condition on absolute accuracy, $F$ (1, 67) = 1.64, $p$ = .204, $\eta_p^2$ = .02 or bias $F$ (1, 67) = 1.87, $p$ = .176, $\eta_p^2$ = .03. The mean bias in the whole sample ($M$ = .18) significantly differed from zero, $t$ (68) = 7.22, $p$ < .001, and thus showed that on average, participants were slightly but significantly overconfident.

The exploratory analysis (Table 2) of incorrect cases only, which would indicate the degree of overconfidence, showed no effect of condition, $F$ (1, 67) = 0.19, $p$ = .665, $\eta_p^2$ < .01. The exploratory analysis of correct cases only, which would indicate the degree of underconfidence, showed a significant effect of condition, $F$ (1, 67) = 6.55, $p$ = .013, $\eta_p^2$ = .09, with the feedback condition being more underconfident ($M$ = -.35) than the no-feedback condition ($M$ = -.25).

Table 2 - Post hoc analysis of confidence and calibration, split up for the cases that were diagnosed correctly or incorrectly.

| | No-feedback condition (n = 35) | | Feedback condition (n = 34) | | Total (n = 69) | |
|---|---|---|---|---|---|---|
| | Mean | SD | Mean | SD | Mean | SD |
| **Incorrect cases** (n = 473) | | | | | | |
| Confidence rating | 5.30 | 1.07 | 5.11 | .86 | 5.20 | .97 |
| Bias | .54 | .19 | .52 | .16 | .53 | .18 |
| **Correct cases** (n = 341) | | | | | | |
| Confidence rating | 6.49 | .80 | 5.90 | 1.03 | 6.20 | .96 |
| Bias | -.25 | .15 | -.35 | .19 | -.30 | .17 |

*Note.* The number of correct or incorrect cases on which the means are based differs for each participant, depending on their performance.

1 Additionally, we analysed only the last nine cases taken together, to give residents the first three cases to learn from the feedback, as did Nederhand et al. (2018). The results did not differ from the analysis of all 12 cases on absolute accuracy, $F$ (1, 67) = 0.90, $p$ = .348, $\eta_p^2$ = .01 or bias $F$ (1, 67) = 2.40, $p$ = .126, $\eta_p^2$ = .04.

## DISCUSSION

It is important for physicians to be able to correctly estimate their diagnostic performance, as overconfidence in a wrong diagnosis might result in diagnostic error and underconfidence in a correct diagnosis may lead to overtesting. The aim of the current study was to investigate whether providing feedback (in the form of the correct diagnosis for a case), would improve diagnostic calibration for more difficult clinical cases. Against expectations, feedback did not improve diagnostic calibration when compared to the control condition without feedback. Exploratory analyses even showed that the feedback made participants significantly more underconfident on correctly diagnosed cases than participants in the control condition.

This finding is at odds with a recent study in which the same type of feedback was shown to improve diagnostic calibration on relatively easy cases (Nederhand et al., 2018). However, we had different cases and a different study population. Also, they had a learning phase of three cases, that we did not include, but when we analysed only the last nine cases, leaving the first three cases to learn from the feedback, the results did not significantly differ from those that we reported. Therefore, there may be two explanations why participants in the feedback condition did not profit from seeing the correct answers for the cases and even became underconfident on correctly diagnosed cases. The first explanation is, that as we used more difficult cases, participants in the feedback group were confronted with their mistakes on some cases, and this may have made them more cautious on subsequent cases, resulting in lower confidence ratings regardless of their actual performance. This fits with an explanation proposed by Raaijmakers et al. (2019), who found, similar to our study, that feedback did not help to improve calibration of future self-assessments.

In the study by Nederhand et al. (2018), in which feedback did improve diagnostic calibration, diagnostic accuracy was very high which suggests that all case were easy. Thus, participants in that study might also simply have adjusted their confidence ratings according to their previous performance and stuck with that rating without considering their actual performance on the present case. Given that they were very likely to give a correct diagnosis, this would lead to higher calibration accuracy. This interpretation also fits with findings from studies in which the difficulty of the cases (Meyer et al., 2013) (or items: Schraw et al., 1993) does vary, but the confidence ratings do not seem to change according to case difficulty and are rather constant (Hacker & Bol, 2019).

A second explanation for why participants did not benefit from the feedback is that the type of feedback we used, may not be helpful for residents to learn how to judge their own performance. Previously it has been found that simple right/wrong feedback has only limited

benefits for improving learning (Ryan et al., 2020). Giving students more elaborate feedback on their performance, that explains why certain answers are right or wrong and the underlying concepts, is more effective for improving performance on future tests. The same may be true for improving future calibration. A review by de Bruin et al. (2017) discusses how physicians (in training) may use *predictive cues* to assess their own performance. In order to judge one's performance, people implicitly make use of a variety of cues (Koriat, 1997). Predictive cues are those cues which help to accurately predict performance, for example when medical experts slow down in clinical practice, they use this as a cue for their difficulty with a case (Moulton et al., 2007). In order for feedback to improve diagnostic calibration, the feedback would need to help physicians to access those predictive cues. We do not yet know what effective predictive cues are for estimating diagnostic performance for physicians in training (de Bruin et al., 2017). However, it has been suggested that providing detailed criteria to judge one's performance can help improve calibration accuracy (Dunlosky et al., 2011; Hawthorne et al., 2017). In our study, participants only got feedback on the end result, which is the diagnosis, and not on the diagnostic process. Providing a performance standard on both the diagnostic process and the correct diagnosis, could possibly help to not only increase their clinical competence, but also to identify cues in the diagnostic process that help them estimate their performance. Future studies should investigate what possible predictive cues are for physicians in training and whether more elaborate feedback would improve diagnostic calibration.



Our study provides new insides into the effect of feedback on diagnostic calibration, but it also has some limitations that should be considered when interpreting the results. First, the study was conducted with fictive, written cases and the residents' performance had no further consequences. The results may have differed in a high-stakes context (Hacker & Bol, 2019), for example in medical practice with real patients, when the task is more important for the residents than it is in an experimental setting. Second, we asked participants to choose only one most likely diagnosis and it could be that, if participants gave an incorrect answer, they had the correct diagnosis in mind as a second or third differential diagnosis. This may also contribute to their tendency to be (slightly) overconfident on average. Third, the way participants had to rate their confidence gives us only limited information on their thought processes and behaviours in clinical practice. Future studies could use different descriptors of confidence, similar to Tweed et al. (2020), by asking participants whether they need more knowledge or information to make a decision, would like to consult a colleague, or feel confident to make a decision on their own. These options may also help to teach physicians in training that seeking help is a valid an valuable option, too (although also in this case, being well-calibrated would help to avoid unnecessary help-seeking). Fourth, we only tested general practice residents and we do not know whether the results apply to physicians with more or less experience or physicians from other disciplines, which may

also contribute to the different results as compared to Nederhand et al. (2018). Fifth, our study does not give us any information on the sources of miscalibration in physicians in training. Future research could focus on this topic, as it may help to find ways to improve diagnostic calibration.

While our study focussed only on (improvement of) diagnostic calibration, future studies could include an estimation of the medical implications that would results from incorrect diagnoses or inadequate confidence. For instance, in the study by Tweed et al. (2017) participants were asked to answer multiple-choice questions on medical cases and rate their certainty. The answers were scored for their level of safeness. They found that when participants were confident about their answer, their response was likely to be either correct or a response that was not causing any patient harm. However, when a participant gave an incorrect answer, the response was more likely to be unsafe when the participant was very confident about it, resulting in a potentially harmful situation for the patient. Helping physicians to better estimate their performance would be especially important for these situations.

To conclude, addressing how we can improve diagnostic calibration is crucial in order to avoid errors (Meyer & Singh, 2019; Zwaan & Hautz, 2019), but proves to be a complex endeavour. It seems unlikely from our results that providing only feedback on the correct diagnosis for a case, will help physicians to better estimate their diagnostic performance; in fact, we found it can even make them less confident about correct diagnoses. This does not mean, however, that feedback cannot have an important role as an educational tool or in medical practice. Paired with a more elaborate intervention that provides participants with cues that are predictive of their actual performance and include safety implications/ harm , it might still be a helpful tool for learning from mistakes (Meyer et al., 2021; Omron et al., 2018; Schiff, 2008). Future studies should investigate whether such more elaborate feedback interventions would be more effective to improve diagnostic calibration.

## REFERENCES


Berner, E. S., & Graber, M. L. (2008). Overconfidence as a Cause of Diagnostic Error in Medicine. *The American Journal of Medicine*, *121*(5 Supplement), S2-S23. https://doi.org/10.1016/j.amjmed.2008.01.001

Cicchetti, D. V. (1994). Guidelines, Criteria, and Rules of Thumb for Evaluating Normed and Standardized Assessment Instruments in Psychology. *Psychological Assessment*, *6*(4), 284-290.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2 ed.). Lawrence Erlbaum Associates.

Costa Filho, G. B., Moura, A. S., Brandão, P. R., Schmidt, H. G., & Mamede, S. (2019). Effects of deliberate reflection on diagnostic accuracy, confidence and diagnostic calibration in dermatology. *Perspectives on Medical Education*, *8*(4), 230-236. https://doi.org/10.1007/s40037-019-0522-5

Davis, D. A., Mazmanian, P. E., Fordis, M., Van Harrison, R., Thorpe, K. E., & Perrier, L. (2006). Accuracy of Physician Self-assessment Compared With Observed Measures of CompetenceA Systematic Review. *JAMA*, *296*(9), 1094-1102. https://doi.org/10.1001/jama.296.9.1094

de Bruin, A., Dunlosky, J., & Cavalcanti, R. (2017). Monitoring and regulation of learning in medical education: The need for predictive cues. *Medical Education*, *51*. https://doi.org/10.1111/medu.13267

Dunlosky, J., Hartwig, M. K., Rawson, K. A., & Lipko, A. R. (2011). Improving college students' evaluation of text learning using idea-unit standards. *The Quarterly Journal of Experimental Psychology*, *64*(3), 467-484. https://doi.org/10.1080/17470218.2010.502239

Eva, K. W., & Regehr, G. (2005). Self-Assessment in the Health Professions: A Reformulation and Research Agenda. *Academic Medicine*, *80*(10), S46-S54. https://journals.lww.com/academicmedicine/Fulltext/2005/10001/Self_Assessment_in_the_Health_Professions__A.15.aspx

Friedman, C. P., Gatti, G. G., Franz, T. M., Murphy, G. C., Wolf, F. M., Heckerling, P. S., Fine, P. L., Miller, T. M., & Elstein, A. S. (2005). Do Physicians Know When Their Diagnoses Are Correct? Implications for Decision Support and Error Reduction. *Journal of General Internal Medicine*, *20*(4), 334-339. https://doi.org/10.1111/j.1525-1497.2005.30145.x

Graber, M. L., Franklin, N., & Gordon, R. (2005). Diagnostic Error in Internal Medicine. *Archives of Internal Medicine*, *165*(13), 1493-1499. https://doi.org/10.1001/archinte.165.13.1493

Hacker, D. J., & Bol, L. (2019). Calibration and Self-Regulated Learning Making the Connections. In *The Cambridge Handbook of Cognition and Education* (pp. 647-677). Cambridge University Press. https://doi.org/10.1017/9781108235631.026

Hacker, D. J., Bol, L., & Bahbahani, K. (2008). Explaining calibration accuracy in classroom contexts: the effects of incentives, reflection, and explanatory style. *Metacognition and Learning*, *3*(2), 101-121. https://doi.org/10.1007/s11409-008-9021-5

Hawthorne, K. A., Bol, L., & Pribesh, S. (2017). Can Providing Rubrics for Writing Tasks Improve Developing Writers' Calibration Accuracy? *The Journal of Experimental Education*, *85*(4), 689-

708. https://doi.org/10.1080/00220973.2017.1299081

Kahi, C. J., Ballard, D., Shah, A. S., Mears, R., & Johnson, C. S. (2013). Impact of a quarterly report card on colonoscopy quality measures. *Gastrointestinal Endoscopy*, *77*(6), 925-931. https://doi.org/10.1016/j.gie.2013.01.012

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General*, *126*(4), 349-370. https://doi.org/10.1037/0096-3445.126.4.349

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 Professions Education*, *6*(4), 506-515. https://doi.org/https://doi.org/10.1016/j.hpe.2020.07.004

Labuhn, A. S., Zimmerman, B. J., & Hasselhorn, M. (2010). Enhancing students' self-regulation and mathematics performance: the influence of feedback and self-evaluative standards. *Metacognition and Learning*, *5*(2), 173-194. https://doi.org/10.1007/s11409-010-9056-2

Lipko, A. R., Dunlosky, J., Hartwig, M. K., Rawson, K. A., Swan, K., & Cook, D. (2009). Using Standards to Improve Middle School Students' Accuracy at Evaluating the Quality of Their Recall. *Journal of Experimental Psychology: Applied*, *15*(4), 307-318.

Meyer, A. N. D., Payne, V. L., Meeks, D. W., Rao, R., & Singh, H. (2013). Physicians' diagnostic accuracy, confidence, and resource requests: a vignette study. *JAMA Internal Medicine*, *173*(21), 1952-1958. https://doi.org/10.1001/jamainternmed.2013.10081

Meyer, A. N. D., & Singh, H. (2019). The Path to Diagnostic Excellence Includes Feedback to Calibrate How Clinicians Think. *JAMA*, *321*(8), 737-738. https://doi.org/10.1001/jama.2019.0113

Meyer, A. N. D., Upadhyay, D. K., Collins, C. A., Fitzpatrick, M. H., Kobylinski, M., Bansal, A. B., Torretti, D., & Singh, H. (2021). A Program to Provide Clinicians with Feedback on Their Diagnostic Performance in a Learning Health System. *The Joint Commission Journal on Quality and Patient Safety*, *47*(2), 120-126. https://doi.org/https://doi.org/10.1016/j.jcjq.2020.08.014

Moulton, C. A., Regehr, G., Mylopoulos, M., & MacRae, H. M. (2007). Slowing down when you should: a new model of expert judgment. *Academic Medicine*, *82*(10 Suppl), S109-116.

Nederhand, M. L., Tabbers, H. K., & Rikers, R. M. J. P. (2019). Learning to calibrate: Providing standards to improve calibration accuracy for different performance levels. *Applied Cognitive Psychology*, *33*(6), 1068-1079. https://doi.org/10.1002/acp.3548

Nederhand, M. L., Tabbers, H. K., Splinter, T. A. W., & Rikers, R. M. J. P. (2018). The Effect of Performance Standards and Medical Experience on Diagnostic Calibration Accuracy. *Health Professions Education*, *4*(4), 300-307. https://doi.org/10.1016/j.hpe.2017.12.008

Omron, R., Kotwal, S., Garibaldi, B. T., & Newman-Toker, D. E. (2018). The Diagnostic Performance Feedback "Calibration Gap": Why Clinical Experience Alone Is Not Enough to Prevent Serious Diagnostic Errors. *AEM education and training*, *2*(4), 339-342. https://doi.org/10.1002/aet2.10119

Raaijmakers, S. F., Baars, M., Paas, F., Van Merriënboer, J. J. G., & Van Gog, T. (2019). Effects of

self-assessment feedback on self-assessment and task-selection accuracy [journal article]. *Metacognition and Learning*, *14*(1), 21-42. https://doi.org/10.1007/s11409-019-09189-5

Rawson, K. A., & Dunlosky, J. (2007). Improving students' self-evaluation of learning of key concepts in textbook materials [doi:10.1080/09541440701326022]. *European Journal of Cognitive Psychology*, *19*(4-5), 559-579. https://doi.org/10.1080/09541440701326022

Ryan, A., Judd, T., Swanson, D., Larsen, D. P., Elliott, S., Tzanetos, K., & Kulasegaram, K. (2020). Beyond right or wrong: More effective feedback for formative multiple-choice tests. *Perspectives on Medical Education*, *9*(5), 307-313. https://doi.org/10.1007/s40037-020-00606-z

Schiff, G. D. (2008). Minimizing Diagnostic Error: The Importance of Follow-up and Feedback. *The American Journal of Medicine*, *121*(5), S38-S42. https://doi.org/https://doi.org/10.1016/j.amjmed.2008.02.004

Schraw, G., Potenza, M. T., & Nebelsick-Gullet, L. (1993). Constraints on the Calibration of Performance. *Contemporary Educational Psychology*, *18*(4), 455-463. https://doi.org/https://doi.org/10.1006/ceps.1993.1034

Shahian, D. M., Normand, S.-L., Torchiana, D. F., Lewis, S. M., Pastore, J. O., Kuntz, R. E., & Dreyer, P. I. (2001). Cardiac surgery report cards: comprehensive review and statistical critique. *The Annals of Thoracic Surgery*, *72*(6), 2155-2168. https://doi.org/https://doi.org/10.1016/S0003-4975(01)03222-2

Tweed, M., Purdie, G., & Wilkinson, T. (2020). Defining and tracking medical student self-monitoring using multiple-choice question item certainty. *BMC Medical Education*, *20*(1), 344. https://doi.org/10.1186/s12909-020-02250-x

Tweed, M. J., Stein, S., Wilkinson, T. J., Purdie, G., & Smith, J. (2017). Certainty and safe consequence responses provide additional information from multiple choice question assessments. *BMC Medical Education*, *17*(1), 1-11.

Zimmerman, B. J. (2008). Investigating Self-Regulation and Motivation: Historical Background, Methodological Developments, and Future Prospects. *American Educational Research Journal*, *45*(1), 166-183. https://doi.org/10.3102/0002831207312909

Zwaan, L., & Hautz, W. E. (2019). Bridging the gap between uncertainty, confidence and diagnostic accuracy: calibration is key. *BMJ Quality & Safety*, *28*(5), 352-355.

## APPENDIX A

*Overview of the chief symptoms and medical conditions that were described in the 12 cases.*

| Chief symptom | Correct diagnosis |
|---|---|
| Diarrhoea | Chronic pancreatitis |
| Shortness of breath | Heart failure |
| Palpitation | Panic disorder |
| Turn dizziness | Benign Paroxysmal Position Vertigo |
| Rash / eczema | Scarlet fever |
| Lower back pain | Spondylodiscitis |
| Amenorrhea | Pregnancy |
| Pain in legs | Spinal canal stenosis |
| Tremor in hand | Multiple sclerosis |
| Facial paralysis | Bell's palsy |
| Rash in the face | Rosacea |
| Vaginal discharge | Bacterial vaginosis |

## APPENDIX B

*Demographics and prior experience ratings.*

| | No-feedback condition | Feedback condition | Total |
|---|---|---|---|
| Sample size | 35 | 34 | 69 |
| Gender | 27 female | 27 female | 54 female |
| Age, *mean* (SD) | 29.23 (2.31) | 29.35 (2.73) | 29.29 (2.51) |
| Prior experience with diagnoses, *mean* (SD) | 2.38 (.52) | 2.43 (.61) | 2.41 (.57) |
| Prior experience with symptoms, *mean* (SD) | 3.21 (.55) | 3.24 (.64) | 3.22 (.59) |

*Note.* Prior experience was rated 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).

# CHAPTER 6

## General Discussion

The first aim of the research presented in this dissertation (i.e., Chapters 2-4) was to test whether we could teach deliberate reflection to physicians in training so that they would apply it when diagnosing new cases without being instructed how to reason. The *deliberate reflection* procedure is a structured way of analysing medical cases (Mamede, Schmidt, & Penaforte, 2008). Its goal is to encourage physicians to check the grounds of their initial diagnostic hypothesis and to consider diagnostic alternatives to avoid mistakes. With deliberate reflection, physicians are asked to systematically list findings that speak in favour of and against their diagnosis, and do this for several possible diagnoses before coming to a final decision. In studies in which physicians were asked to follow the deliberate reflection procedure, it showed to be effective for improving diagnostic accuracy when cases were difficult (Mamede, Schmidt, & Penaforte, 2008), or when physicians were susceptible to cognitive bias (Mamede, Splinter, et al., 2012; Mamede et al., 2010; Schmidt et al., 2014). In these studies, physicians were actively instructed to follow the procedure. For deliberate reflection to help prevent error in clinical practice more effectively, physicians would need to learn the steps and then apply them rapidly and autonomously when diagnosing a new case.

The second aim of the research presented in this dissertation (i.e., Chapter 5) was to see whether we could help physicians in training to improve their diagnostic calibration (which would help them to determine, for instance, when they would need to apply further reflection) by providing them with performance feedback. There is some evidence that physicians tend to engage in more reflective reasoning when diagnosing a case they perceive as difficult (Mamede, Schmidt, Rikers, et al., 2008; Noyer et al., 2017). However, they are not always good at recognising the instances in which the case would still need extra attention because they have not yet diagnosed it correctly (Costa Filho et al., 2019; Davis et al., 2006; Meyer et al., 2013). The relationship between physicians' diagnostic performance and their confidence in their diagnosis is called *diagnostic calibration.* Improving physicians' diagnostic calibration may on the one hand help to prevent premature closure and the resulting diagnostic errors, meaning that they do not accept a diagnosis before it is sufficiently verified (Berner & Graber, 2008; McSherry, 1997), and it may help physicians to recognise in which topics they would benefit from further improving their knowledge (Zimmerman, 2008). On the other hand, it may also help physicians to recognise when their diagnosis *is* correct and prevent unnecessary further diagnostic testing for the patients. We conducted all studies in the context of general practice, with cases that described consultations in Dutch general practices. General practitioners are often the first to see a patient and have to decide whether treatment or referral to a different specialist is required. This makes their judgement, and therefore good diagnostic reasoning skills, crucial for a patient's trajectory.

In this chapter, I will first summarize the studies' main findings, followed by a discussion of (potential explanations for) the main findings, implications for educational practice, challenges and limitations of the studies, and ideas for future research.

## Summary of the Main Findings

The studies described in Chapter 2, 3 and 4 focussed on the first aim and tested whether deliberate reflection could be taught to physicians in training so that they would apply it when diagnosing novel cases in a later test. The first two studies focussed on residents in the general-practice training and the third study on medical students.

The study in **Chapter 2** tested the hypothesis that learning deliberate reflection via *example-based learning* would be more effective than via *learning-by-doing*, and that both would be more effective (i.e., residents would show more elements of deliberate reflection in their reasoning when diagnosing future novel cases) compared to a control condition in which students were not taught deliberate reflection. The study consisted of a learning session and two test sessions. In the learning session, the residents diagnosed cases either by following the steps of deliberate reflection themselves (learning–by-doing; $n$ = 11), or by studying examples of how an expert had diagnosed the case with deliberate reflection (example-based learning; $n$ = 19), or without any instructions on how to reason (control; $n$ = 14). In two tests, a same-day test a couple of hours later and a delayed test a week later, they diagnosed 12 novel cases per test, four of which were unrelated to the cases in the learning session and eight were related to them. During both tests, we measured whether the intervention had an effect on diagnostic accuracy for novel cases. In the delayed test, we also evaluated whether participants used deliberate reflection when diagnosing future cases. As an indication of the residents' reasoning process we used a justification task *after* diagnosing, in which they had to explain how they had arrived at their diagnosis. The results showed that, against expectations, there was no significant difference between learning-by-doing, example-based learning and the control condition for either the related or unrelated novel cases. Participants who were taught deliberate reflection did not show more elements of the reflection procedure in their reasoning when diagnosing future test cases compared to participants in the control condition, nor did their diagnostic accuracy improve.

From the results of this study, we could not infer whether the aim of making residents apply deliberate reflection autonomously, could not be reached, or whether they had not learned the procedure properly. For example, it may have been the case that the residents in the learning by doing and example-based learning conditions focussed more on the content of the cases than on the deliberate reflection procedure used for solving them. Also, the

sample in the study was small, making it difficult to draw conclusions. For this reason, we used a slightly different approach in the next study.

In the study described in **Chapter 3**, we used example-based learning combined with *learning-by-teaching* as an intervention to teach deliberate reflection. This intervention had several advantages: By asking participants to teach the steps of deliberate reflection to a fictitious peer their attention would be directed towards the procedure itself during example study, and by recording their explanations we would be able to see whether they had learned the procedure and were able to explain it correctly. In this study we tested the hypothesis that residents who had learned deliberate reflection via learning-by-teaching would show more elements of deliberate reflection in their reasoning when diagnosing future novel cases than residents in a control condition who were not taught deliberate reflection. The study consisted of a learning session and a delayed test session. In the learning session, residents either studied examples of deliberate reflection and then explained the procedure and how it was applied while being video recorded (learning-by-teaching; $n = 28$) or they diagnosed cases without further instructions (control; $n = 28$). During the test session, all participants diagnosed four new cases while thinking aloud. The think-aloud protocols were analysed to evaluate the residents' reasoning process. The recordings of the explanations from the learning session showed that participants in the deliberate reflection condition had learned the procedure and were able to explain it correctly. However, we again did not find that residents in this condition showed more elements of deliberate reflection in their think-aloud protocols of the test cases than residents in the control condition, nor were there significant differences in diagnostic accuracy.

Thus, residents did not seem to apply deliberate reflection on future cases, even though it was clear now that they had learned the procedure. This may have been the case because the residents were already too experienced in diagnosing cases and therefore it may have been difficult to change their reasoning routine. If this was the case, one should find benefits of teaching deliberate reflection for less experienced participants, like medical students. Another reason why reflection was not applied, might have been that the need to do so was not felt on the test cases, because they did not feel that the cases were that difficult.

Therefore, the study described in **Chapter 4** tested whether medical students would learn deliberate reflection via a combination of example-based learning and learning-by-teaching, but would only apply it on future cases when they thought that cases would be difficult and therefore required extra attention. The study consisted of a learning session and a delayed test session. In the learning session, students were randomly assigned to the deliberate reflection condition ($n = 58$) with learning-by-teaching, or the control condition ($n = 61$), following the same procedure as in the study described in Chapter 3. In the test session,

students diagnosed six ambiguous cases, followed by a recall task after diagnosing, in which they were asked to recall and write down everything they remembered from the case. The goal of the recall task was to measure on which parts of the case the participants had focussed. As this was done after having diagnosed the case, we expected that it would not have an influence on the diagnostic process as the measurements in the previous studies might have. After having diagnosed half of the cases, they were told that the next cases would be difficult ones. The results indicated that this manipulation had the expected effect as students reported more mental effort when solving a case and were less confident about their diagnosis when a case was described as difficult. The recall task was analysed for an indication of the students' reasoning process. We counted whether participants recalled the important features for the case, which helped to discriminate between the possible diagnoses, and whether these recalled features were related to their own diagnosis or the alternative diagnosis.

The results showed that the students who had been taught deliberate reflection recalled more features for the *alternative* diagnosis than the control condition, regardless of the type of case (described as difficult or not). This may indicate that they used key elements of the deliberate reflection procedure, instead of focussing on their own diagnosis only. Against expectations, they did also apply it when cases were not described as difficult, meaning that they did not need a trigger to apply deliberate reflection. The deliberate reflection condition also recalled more features related to their own diagnosis on the first three cases than did the control condition, but on the last three cases (described as difficult), the differences were no longer significant. Students in the control condition recalled more of the important features of the case (for their own as well as the alternative diagnosis) when cases were described as difficult compared to the cases that were not described as difficult. This indicates that they engaged in more reflective reasoning when thinking that they would encounter a difficult case triggered them to do so.

Finally, the study described in **Chapter 5** focussed on the second aim of this dissertation and tested the hypothesis that feedback on previous diagnostic performance would improve physicians' diagnostic calibration on future cases. In this study, general-practice residents diagnosed 12 written cases. The residents were assigned to either the feedback ($n$ = 34) or the no-feedback condition ($n$ = 35). In the feedback condition, participants diagnosed a case, rated how confident they were about their given diagnosis. Then, they received the correct diagnosis for the case and were asked to compare that diagnosis to the one they had given. Participants in the no-feedback condition followed the same steps but without receiving the correct diagnosis. From the participants' diagnostic accuracy and confidence ratings we calculated calibration in terms of *absolute accuracy* and *bias*. Absolute accuracy describes the difference between diagnostic accuracy and confidence, and bias de-

scribes the direction of that difference (over- or underconfidence). The results showed that feedback did not improve calibration, as no significant difference was found between the feedback and no-feedback conditions on either of the two calibration measures. Two post hoc analyses were conducted to better understand the results. One analysed calibration only in instances in which a resident had given a correct diagnosis. The other analysed calibration only in instances in which a resident had given an incorrect diagnosis. It was found that participants in the feedback condition were significantly more underconfident for cases they had diagnosed correctly than participants in the no-feedback condition. There was no difference between conditions on incorrectly diagnosed cases.

## Discussion of the Main Findings

**Learning and applying deliberate reflection**. In the first two studies in residents on learning deliberate reflection (Chapter 2 & 3), the interventions were not effective to change future reasoning, even though the results of the second study (Chapter 3) indicated that participants had learned the deliberate reflection procedure. In the third study (Chapter 4). However, medical students who had been taught deliberate reflection did seem to apply key elements of it when diagnosing future cases, even when they were not told that cases would be difficult. The results indicate that deliberate reflection can be learned and that some elements of it may then be applied to novel cases with a different medical content. Students in the control condition also seemed to apply more reflective reasoning when they thought that the cases would be difficult ones, but students in the deliberate-reflection condition did not need this trigger.

There are several possible explanations why no effect was found in the first two studies, but in the third study participants seemed to adopt deliberate reflection. This may be the case because the students in the third study were less experienced with diagnosing cases than the residents in the first two studies. The different findings may also be the result of the different methods used to measure the participants' reasoning. The recall task used with medical students focussed on different aspects of the reasoning process than the justification or think-aloud task used with residents. When analysing the results of the justification and think-aloud task, it was counted how many alternative diagnoses to their first diagnosis they considered and how many (absent) symptoms contradicting a diagnosis, as we assumed that these are the elements of deliberate reflection that help to challenge one's first impression of the case and avoid a confirmation bias. With the recall task, we used a different approach. The cases used for the recall task all had two diagnoses that stood out as being very likely for the case. It was measured how many features students could recall that were related to either the diagnosis they had given or the alternative diagnosis. We assumed that they would recall more features if they had focussed on them during the

diagnostic task (Long et al.). This way it was measured whether the students paid attention to the details of the case as an indication of conscious thought (Mamede et al., 2007), and whether they had focussed not only on their own but also on the alternative diagnosis. In contrast to the analysis of the justification and think-aloud task, it was not analysed whether they specifically considered symptoms contradicting their diagnosis and symptoms that were absent in the case. It may either be that these specific elements do not change from learning deliberate reflection, or that the justification and think-aloud task influenced all participants' reasoning (including the control group) so that they engage in more reflective reasoning, which means that these tasks would not give a good indication of their usual diagnostic reasoning. While Chapter 3 showed that participants learned deliberate reflection, the recall task in Chapter 4 can only give indirect evidence of them applying it. However, the recall task may be a better method to describe on which parts of the case the participants focussed during diagnostic reasoning than the justification or think-aloud task.

Another explanation why we only found an effect in the third study could be that residents are already too experienced with an internalised own way of diagnostic reasoning. This makes it more difficult to change this routine with a short intervention, whereas medical students are still learning how to perform diagnostic reasoning and therefore may adopt a new technique more easily. Between the learning and the (delayed) test session, the general-practice residents were working in medical practice where they were also seeing and diagnosing their own patients without instructions on how to reason. This also may have diminished the effect that learning deliberate reflection may have had on their reasoning. Therefore, it may be the most efficient to teach deliberate reflection early on during medical education. When physicians already have more experience, they may need a more extensive learning intervention to adopt a new procedure, for example by seeing their supervisors in general practice implement it, but more studies are necessary to test this. The effects of learning deliberate reflection would, however, be expected to be more visible with residents then with students, as physicians are prone to using more pattern recognition and non-analytical reasoning as they gain more experience (Elstein & Schwartz, 2002; Schmidt & Boshuizen, 1993; Schmidt et al., 1990). That we were able to detect an effect with students in Chapter 4 may show that even students rely on pattern recognition to some degree, as has been found by Tay et al. (2016).

It may, however, also be the case that the residents' internalised own way of diagnostic reasoning resembles the deliberate reflection method and therefore we did not find an effect. While the residents had not been taught deliberate reflection before, they do have clinical reasoning lessons in which they are taught to analyse several differential diagnoses for a case. Although they do this in a less structured way, deliberate reflection may not be very different from what they already learn during clinical reasoning lessons for an effect

to be found in the residents' diagnostic reasoning or for a benefit of diagnostic accuracy to occur.

Another potential explanation for the findings is that the participants in the different studies may have had a different perception of the difficulty of the test cases. Physicians have been found to engage in more reflective reasoning when they perceive a case as being difficult (Mamede, Schmidt, Rikers, et al., 2008; Noyer et al., 2017). Therefore, it may be that they will only apply the learned deliberate-reflection procedure if they feel that a case requires that extra attention. While students showed that they did not need a trigger to apply deliberate reflection after being taught the procedure, this may be different for residents. It may be that residents did not feel the need to apply deliberate reflection in the test cases, for example because they did not think that they were difficult enough to need that extra attention. The medical students on the other hand may have perceived all cases as being relatively difficult, even when their perception of difficulty increased more when they were told that the upcoming cases would be difficult ones. With a more extensive learning phase and cases that are perceived as more difficult, residents may also show signs of reflective reasoning in a recall task. This could be tested in future studies.

**Improving diagnostic calibration**. As for the last study (Chapter 5), on the effect of feedback on diagnostic calibration, our results did not show an improvement in calibration. This was unexpected, because findings from a similar previous study by (Nederhand et al., 2018) showed that performance feedback did improve calibration. In our study, the performance feedback even made the residents more underconfident about cases they had actually diagnosed correctly. One potential explanation for why we did not replicate the results by Nederhand et al., might lie in the case difficulty. While in the study by Nederhand et al. diagnostic accuracy was high, in our study it was lower, indicating that the cases were more difficult to solve. Therefore, most residents diagnosed some cases correctly and others incorrectly, and the feedback confronted them with these errors. As a result, participants in the feedback condition seemed to lower their confidence ratings even when they had diagnosed a case correctly, leading to more underconfidence instead of better calibration. In the study by Nederhand et al. participants gave correct diagnoses for most cases. In that study, raising confidence ratings for all cases would have led to an improvement in calibration, but it does not necessarily mean that participants had learned to better differentiate between correct and incorrect diagnoses.

## Challenges and Limitations

One of the challenges we encountered was to find a good measurement of the participants' reasoning process when they were diagnosing cases in the test phase. That measurement

should be able to indicate whether their reasoning follows the steps of deliberate reflection, but should not influence and alter their reasoning. With the justification task (Chapter 2) and the think-aloud task (Chapter 3) we tried to follow along the thought process during diagnosing. The justification task asked participants to reproduce their thought process, while the think-aloud task measured their reasoning at the time of diagnosing a case. The advantage of both methods is, that the results were so detailed that they could be used to reconstruct the deliberate-reflection table and to see whether their reasoning pattern fits within the steps of deliberate reflection. The disadvantage of both methods it, that they may not give a correct picture of how participants would diagnose usually, without being given a think-aloud or justification task. The act of thinking aloud may already trigger more reflective reasoning, but also when we asked participants to reconstruct their though processes afterwards (justification task), they may have engaged in a new, more reflective reasoning process and reported that instead of their original diagnostic reasoning. The recall task (Chapter 4) on the other hand is a more indirect measurement of the reasoning process. The advantage of this is, that it is less likely to have an influence on the diagnostic reasoning process it aims to measure. The disadvantage of the recall task is, that the data is less detailed. It indicates on which part of the case a participant has focussed, but it does not show the process itself and why certain choices were made.

It is important to note that in our studies we only focussed on deliberate reflection, because this intervention has shown to improve diagnostic performance (Mamede, Schmidt, & Penaforte, 2008; Mamede, Splinter, et al., 2012; Mamede et al., 2010; Schmidt et al., 2014). Therefore, we aimed to measure whether participants were applying the critical elements of deliberate reflection. Reflective reasoning in the broader sense, however, does not necessarily follow these steps. Reflection has been defined in many different forms (Ng et al., 2015; Schaepkens et al., 2021), can occur during the diagnostic task but also afterwards and may not even be something that can be measured (de la Croix & Veen, 2018). This means that even if physicians do not follow the specific steps of the deliberate-reflection procedure, they will be engaged in some form of reflection, which we did not measure in our studies.



Another challenge was to make cases that were difficult enough for residents, but where the expert general practitioners could still agree on one correct diagnosis, making it possible to measure diagnostic accuracy. This means that the cases may not represent the whole spectrum of cases they would encounter in general practice, where it may not always be possible to come to a conclusive diagnosis. We do not know what the results would have been with even more difficult cases. It may be that they would have been more likely to engage in deliberate reflection after having learned the procedure and they may have found it more difficult to estimate their diagnostic accuracy (Costa Filho et al., 2019; Meyer et al., 2013).

A final limitation is that the sample sizes were quite small, especially in the first study described in Chapter 2. Furthermore, when we found a statistically significant effect, it often had a small to medium effect size and we do not know whether the effect would still be found after a longer period. Therefore, the results should be interpreted with caution until they are replicated in future studies.

## Implications for Medical Education and Future Research

Our studies show that the combination of example-based learning and learning-by-teaching can be effective for teaching a novel reasoning strategy for diagnostic reasoning, at least for medical students. This is in line with research showing that this method is effective for novices to learn problem-solving skills (Hoogerheide, Renkl, et al., 2019) and reasoning skills (Hoogerheide et al., 2014). When implementing this procedure in medical education, however, it is important to consider some practical aspects. A disadvantage of learning-by-teaching as conducted in our studies is that it required a lot of preparation and material like the video cameras, headphones, and microphones. Some students felt distracted by the other students' talking or felt uncomfortable talking when others could hear them. To apply the method in educational practice, it may be more convenient and still effective to give teaching as an homework assignment (Hoogerheide, Visee, et al., 2019). The students or residents could then record the videos on their phones in an environment where they feel comfortable doing so. Future studies could test whether this would lead to similar results.

As mentioned earlier, a reason why the interventions to teach deliberate reflection were effective in the study described in Chapter 4, but ineffective in the studies described in Chapter 2 and 3, may be that the level of experience of the participants was different in the studies. To test this explanation, future studies could replicate the study design described in Chapter 4 and conduct the study among residents. If the findings would again show that teaching deliberate reflection is only effective with medical students, it would mean that the intervention is not effective enough for residents. For medical education this would mean that in order to teach deliberate reflection to physicians in training, it would be best to start early on in their medical training when they have not yet much experience with diagnosing cases. Medical students seemed to adopt (some elements of) deliberate reflection after only a short learning session. For residents, other interventions or a longer learning phase may be more effective. If, however, a replication study would find an effect of learning deliberate reflection with residents, too, it would mean that the recall task measures different elements of the reasoning process than the justification and think-loud task and only these elements changed, or that the recall task is more sensitive.

As a next step for improving diagnostic reasoning, it would be important to investigate whether the effect found in Chapter 4 would also lead to improvements in diagnostic performance, which would ultimately be the goal of the diagnostic reasoning education. In that study, we tested and found an effect of practicing with deliberate reflection on the students' reasoning when solving novel cases. We did not test, however, what the effect on diagnostic accuracy was, as the test cases were not designed for this. Previous studies found that engaging in the complete steps of deliberate reflection improves diagnostic accuracy in some cases (Mamede, Schmidt, & Penaforte, 2008; Mamede, Splinter, et al., 2012; Mamede et al., 2010; Schmidt et al., 2014). Future studies should test whether this is also the case after having learned deliberate reflection when physicians may only apply some elements instead of following the complete procedure.

Another important direction for future research is to find out which of the steps make deliberate reflection effective for improving diagnostic performance. In the present studies, we taught the complete deliberate reflection steps. For practice, it would be useful to know if physicians would only need to apply a part of deliberate reflection. This would make it more efficient and maybe easier for physicians to apply this during a busy workday with time constraints. For example, generating multiple possible diagnoses early in the diagnostic process may already help to avoid a tunnel vision, as it has been shown that physicians tend to stick with their first impression of a case even if contradicting evidence is presented (Kostopoulou et al., 2012; Kostopoulou et al., 2017). Furthermore, listing findings that speak against the diagnosis at hand as well as findings that would be expected if the diagnosis were true but are absent in the case, may help to critically evaluate the diagnoses and avoid confirmation bias. It would be interesting to test whether listing findings supporting a diagnosis does help with finding a correct diagnosis or whether this step could be left out, as it could be expected that physicians naturally generate diagnoses based on present symptoms that are related to that diagnosis (Gruppen et al., 1988; Pelaccia et al., 2014; Wortman, 1972). Furthermore, in the studies where deliberate reflection proved to be effective, physicians were asked to fill in the reflection table either on the computer or on paper. Visualisations haven been found to be beneficial for decision making in various domains, including medicine (Padilla et al., 2018). Therefore, visually sorting the symptoms and diseases in the table may help with the diagnostic process. Future studies should test whether this is more beneficial for diagnostic performance than only following the steps in one's head or following the steps while thinking aloud, for example when discussion a difficult case with a colleague, which may be less time consuming and thus more efficient than writing.

Another open question that is relevant for medical education, is how long lasting the changes in diagnostic reasoning are, that were found in Chapter 4. In order to make it part of the physicians' reasoning routine, it may be necessary to use deliberate reflection as a regular

and recurrent part of medical education. For example, it could be used during different courses where clinical cases are being discussed, and the students' or residents' supervising physicians could model the procedure in medical practice. Also, as students gain more experience, their reasoning is likely to change towards more reliance on pattern recognition (Elstein & Schwartz, 2002; Schmidt & Boshuizen, 1993; Schmidt et al., 1990). This is very helpful because it makes the reasoning process efficient, but it may also decrease the effect of learning deliberate reflection. Therefore, it may be necessary to teach deliberate reflection more often during medical training before long term effects can be achieved to help physicians correct their initial errors when encountering difficult cases in medical practice.

If future studies would not replicate our findings with the students, but would find that they do not adopt deliberate reflection, it would not mean that deliberate reflection cannot be beneficial for medical education. In addition to learning the procedure itself, deliberate reflection has been shown to be effective for students as a means to learn about the causes, signs, and symptoms of the considered diseases (Mamede, Van Gog, et al., 2012; Mamede et al., 2014). In these studies, students who were asked to analyse clinical cases by following the deliberate-reflection procedure performed better in a subsequent test, diagnosing new cases of the same diseases, than students who followed a more conventional approach to solve them such as giving differential diagnoses. Giving students more instructional guidance during reflection turned out to be even more effective (Fernandes et al., 2021; Ibiapina et al., 2014; Mamede et al., 2019). When students were either provided with the diagnoses on which they should reflect, or when they received a complete modelling example of deliberate reflection on a case, students performed higher on similar cases one week later than students who had engaged in free reflection without further guidance. This means, that deliberate reflection can be used for students who need to learn about different diagnoses even if they do not learn the procedure itself.

In order to help physicians recognise the situations in which more reflection on a case is needed, future studies could test different interventions to improve diagnostic calibration. If physicians recognise cases that are more difficult to solve and need extra attention, this may naturally evoke more reflective reasoning (Mamede, Schmidt, Rikers, et al., 2008; Noyer et al., 2017) as is also suggested by the results described in Chapter 4 where students in the control condition did also engage in more reflective reasoning when they were told that cases would be difficult. Improving diagnostic calibration means that physicians are better at recognising when a case has not been diagnosed correctly yet and, therefore, the diagnosis could still be improved by applying deliberate reflection.

Contrary to Nederhand et al. (2018), our study (Chapter 5) showed that providing physicians in training with performance feedback did not help to improve diagnostic calibration.

Therefore, future studies could test whether different interventions would be more effective. For example, it may help to give more informative feedback to explain why an answer was right or wrong, as is has been suggested that this type of feedback is more effective for helping students with self-assessment and self-regulated learning (Nicol & Macfarlane-Dick, 2006). In addition to that, informative feedback has been found to be more effective than right/wrong feedback to help students improve their medical knowledge (Ryan et al., 2020). It may also help to encourage physicians to pay attention to several measures when making self-assessments, for example how much time it took them to diagnose a case (Eva & Regehr, 2007; Zwaan & Hautz, 2019). Calibration may also be improved by giving physicians incentives to make good self-assessments (Hacker et al., 2008). Some studies suggest that calibration is knowledge-related and could be improved by improving medical knowledge (Costa Filho et al., 2019). Others suggest that it is a general trait (Carpenter et al., 2019) which may mean that overall summative feedback may be more effective than case-by-case feedback. Furthermore, future studies could test whether individual factors help to explain calibration and how to improve it.

Another difficulty with improving calibration is, that even if calibration is improved in an experimental setting, these interventions are often not as effective in practice (Hacker et al., 2008). The same may be true for learning deliberate reflection. All four studies in this dissertation were experimental studies conducted on computers with written cases. This controlled setting allows us to research causal relations. However, the findings of these studies may not always reflect medical practice. Therefore, it would be necessary to conduct studies in a medical practice context with actual patients before any recommendations for improving diagnostic calibration or learning deliberate reflection in medical practice can be given.



## Conclusion

The main purpose of this dissertation was to investigate whether and how physicians in training could learn *deliberate reflection* for diagnosing medical cases, and how their *diagnostic calibration* (i.e., how well physicians can estimate their own diagnostic performance) could be improved. The studies showed that participants did not easily adopt deliberate reflection when diagnosing future novel cases. Only one study found an effect, which may be the case because the participants in that study were more inexperienced with diagnosing cases, or because the task in the test phase was better suited for measuring the changes in their reasoning. For improving diagnostic calibration, we found performance feedback to be ineffective, which was at odds with previous studies. We hope that these studies will inspire future research to better understand under which circumstances deliberate reflection can best be learned, how physicians can be supported to make better estimation of their own diagnostic performance, and whether it has the desired outcomes for patient safety.

## REFERENCES

Berner, E. S., & Graber, M. L. (2008). Overconfidence as a Cause of Diagnostic Error in Medicine. *The American Journal of Medicine*, *121*(5 Supplement), S2-S23. https://doi.org/10.1016/j.amjmed.2008.01.001

Carpenter, J., Sherman, M. T., Kievit, R. A., Seth, A. K., Lau, H., & Fleming, S. M. (2019). Domain-general enhancements of metacognitive ability through adaptive training. *Journal of Experimental Psychology: General*, *148*(1), 51-64.

Costa Filho, G. B., Moura, A. S., Brandão, P. R., Schmidt, H. G., & Mamede, S. (2019). Effects of deliberate reflection on diagnostic accuracy, confidence and diagnostic calibration in dermatology. *Perspectives on Medical Education*, *8*(4), 230-236. https://doi.org/10.1007/s40037-019-0522-5

Davis, D. A., Mazmanian, P. E., Fordis, M., Van Harrison, R., Thorpe, K. E., & Perrier, L. (2006). Accuracy of Physician Self-assessment Compared With Observed Measures of CompetenceA Systematic Review. *JAMA*, *296*(9), 1094-1102. https://doi.org/10.1001/jama.296.9.1094

de la Croix, A., & Veen, M. (2018). The reflective zombie: Problematizing the conceptual framework of reflection in medical education. *Perspectives on Medical Education*, *7*(6), 394-400. https://doi.org/10.1007/s40037-018-0479-9

Elstein, A. S., & Schwartz, A. (2002). Clinical problem solving and diagnostic decision making: selective review of the cognitive literature. *BMJ*, *324*(7339), 729-732. https://doi.org/10.1136/bmj.324.7339.729

Eva, K. W., & Regehr, G. (2007). Knowing when to look it up: a new conception of self-assessment ability. *Academic Medicine*, *82*(10 Suppl), S81-84.

Fernandes, R. A. F., Malloy-Diniz, L. F., de Vasconcellos, M. C., Camargos, P. A. M., & Ibiapina, C. (2021). Adding guidance to deliberate reflection improves medical student's diagnostic accuracy. *Medical Education*, *55*(10), 1161-1171.

Gruppen, L. D., Woolliscroft, J. O., & Wolf, F. M. (1988). The contribution of different components of the clinical encounter in generating and eliminating diagnostic hypotheses. *Proceedings of the Annual Conference on Research in Medical Education Conference on Research in Medical Education*, *27*, 242-247. https://www.scopus.com/inward/record.uri?eid=2-s2.0-0024265676&partnerID=40&md5=ec1a3e3ca97d7384464f53fab1aab0b3

Hacker, D. J., Bol, L., & Bahbahani, K. (2008). Explaining calibration accuracy in classroom contexts: the effects of incentives, reflection, and explanatory style. *Metacognition and Learning*, *3*(2), 101-121. https://doi.org/10.1007/s11409-008-9021-5

Hoogerheide, V., Loyens, S. M. M., & Van Gog, T. (2014). Effects of creating video-based modeling examples on learning and transfer. *Learning and Instruction*, *33*, 108-119. https://doi.org/10.1016/j.learninstruc.2014.04.005

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.

*Journal of Educational Psychology*, *111*(1), 45-56.

Hoogerheide, V., Visee, J., Lachner, A., & Van Gog, T. (2019). Generating an instructional video as homework activity is both effective and enjoyable. *Learning and Instruction*, *64*, 101226. https://doi.org/https://doi.org/10.1016/j.learninstruc.2019.101226

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. *Medical Education*, *48*, 796-805. https://doi.org/10.1111/medu.12435

Kostopoulou, O., Russo, J. E., Keenan, G., Delaney, B. C., & Douiri, A. (2012). Information Distortion in Physicians' Diagnostic Judgments. *Medical Decision Making*, *32*(6), 831-839. https://doi.org/10.1177/0272989x12447241

Kostopoulou, O., Sirota, M., Round, T., Samaranayaka, S., & Delaney, B. C. (2017). The Role of Physicians' First Impressions in the Diagnosis of Possible Cancers without Alarm Symptoms. *Medical decision making : an international journal of the Society for Medical Decision Making*, *37*(1), 9-16. https://doi.org/10.1177/0272989x16644563

Long, N. M., Kuhl, B. A., & Chun, M. M. Memory and Attention. In *Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience* (pp. 1-37). https://doi.org/https://doi.org/10.1002/9781119170174.epcn109

Mamede, S., Figueiredo-Soares, T., Elói Santos, S. M., De Faria, R. M. D., Schmidt, H. G., & Van Gog, T. (2019). Fostering novice students' diagnostic ability: the value of guiding deliberate reflection. *Medical Education*, *53*(6).

Mamede, S., Schmidt, H. G., & Penaforte, J. C. (2008). Effects of reflective practice on the accuracy of medical diagnoses. *Medical Education*, *42*(5), 468-475. https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1365-2923.2008.03030.x

Mamede, S., Schmidt, H. G., Rikers, R. M., Penaforte, J. C., & Coelho-Filho, J. M. (2008). Influence of perceived difficulty of cases on physicians' diagnostic reasoning. *Academic Medicine*, *83*(12), 1210-1216. https://doi.org/10.1097/ACM.0b013e31818c71d7

Mamede, S., Schmidt, H. G., Rikers, R. M. J. P., Penaforte, J. C., & Coelho-Filho, J. M. (2007). Breaking down automaticity: Case ambiguity and the shift to reflective approaches in clinical reasoning. *Medical Education*, *41*(12), 1185-1192. https://doi.org/10.1111/j.1365-2923.2007.02921.x

Mamede, S., Splinter, T. A., Van Gog, T., Rikers, R. M., & Schmidt, H. G. (2012). Exploring the role of salient distracting clinical features in the emergence of diagnostic errors and the mechanisms through which reflection counteracts mistakes. *BMJ Quality & Safety*, *21*(4), 295-300. https://doi.org/10.1136/bmjqs-2011-000518

Mamede, S., Van Gog, T., Moura, A. S., De Faria, R. M., Peixoto, J. M., Rikers, R. M., & Schmidt, H. G. (2012). Reflection as a strategy to foster medical students' acquisition of diagnostic competence. *Medical Education*, *46*(5), 464-472. https://doi.org/10.1111/j.1365-2923.2012.04217.x

Mamede, S., Van Gog, T., Moura, A. S., de Faria, R. M. D., Peixoto, J. M., & Schmidt, H. G. (2014).

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. *Academic Medicine*, *89*, 121-127.

Mamede, S., Van Gog, T., Van den Berge, K., Rikers, R. M., Van Saase, J. L., Van Guldener, C., & Schmidt, H. G. (2010). Effect of Availability Bias and Reflective Reasoning on Diagnostic Accuracy Among InternalMedicine Residents. *JAMA*, *304*(11), 1198-1203. https://doi.org/10.1001/jama.2010.1276

McSherry, D. (1997). Avoiding premature closure in sequential diagnosis. *Artificial Intelligence in Medicine*, *10*(3), 269-283. https://doi.org/https://doi.org/10.1016/S0933-3657(97)00396-5

Meyer, A. N. D., Payne, V. L., Meeks, D. W., Rao, R., & Singh, H. (2013). Physicians' diagnostic accuracy, confidence, and resource requests: a vignette study. *JAMA Internal Medicine*, *173*(21), 1952-1958. https://doi.org/10.1001/jamainternmed.2013.10081

Nederhand, M. L., Tabbers, H. K., Splinter, T. A. W., & Rikers, R. M. J. P. (2018). The Effect of Performance Standards and Medical Experience on Diagnostic Calibration Accuracy. *Health Professions Education*, *4*(4), 300-307. https://doi.org/10.1016/j.hpe.2017.12.008

Ng, S. L., Kinsella, E. A., Friesen, F., & Hodges, B. (2015). Reclaiming a theoretical orientation to reflection in medical education research: a critical narrative review. *Medical Education*, *49*(5), 461-475.

Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: a model and seven principles of good feedback practice [doi: 10.1080/03075070600572090]. *Studies in Higher Education*, *31*(2), 199-218. https://doi.org/10.1080/03075070600572090

Noyer, A. L., Esteves, J. E., & Thomson, O. P. (2017). Influence of perceived difficulty of cases on student osteopaths' diagnostic reasoning: a cross sectional study. *Chiropractic & Manual Therapies*, *25*, 32-32. https://doi.org/10.1186/s12998-017-0161-z

Padilla, L. M., Creem-Regehr, S. H., Hegarty, M., & Stefanucci, J. K. (2018). Decision making with visualizations: a cognitive framework across disciplines. *Cognitive Research: Principles and Implications*, *3*(1), 29. https://doi.org/10.1186/s41235-018-0120-9

Pelaccia, T., Tardif, J., Triby, E., Ammirati, C., Bertrand, C., Dory, V., & Charlin, B. (2014). How and When Do Expert Emergency Physicians Generate and Evaluate Diagnostic Hypotheses? A Qualitative Study Using Head-Mounted Video Cued-Recall Interviews. *Annals of Emergency Medicine*, *64*(6), 575-585. https://doi.org/https://doi.org/10.1016/j.annemergmed.2014.05.003

Ryan, A., Judd, T., Swanson, D., Larsen, D. P., Elliott, S., Tzanetos, K., & Kulasegaram, K. (2020). Beyond right or wrong: More effective feedback for formative multiple-choice tests. *Perspectives on Medical Education*, *9*(5), 307-313. https://doi.org/10.1007/s40037-020-00606-z

Schaepkens, S. P. C., Veen, M., & de la Croix, A. (2021). Is reflection like soap? a critical narrative umbrella review of approaches to reflection in medical education research. *Advances in Health Sciences Education*. https://doi.org/10.1007/s10459-021-10082-7

Schmidt, H. G., & Boshuizen, H. P. A. (1993). On Acquiring Expertise in Medicine. *Educational*

*Psychology Review*, *5*(3), 205-221.

Schmidt, H. G., Mamede, S. I., Van Den Berge, K., Van Gog, T., Van Saase, J. L. C. M., & Rikers, R. M. J. P. (2014). Exposure to media information about a disease can cause doctors to misdiagnose similar-looking clinical cases. *Academic Medicine*, *89*(2).

Schmidt, H. G., Norman, G., & Boshuizen, H. (1990). A cognitive perspective on medical expertise: theory and implications. *Academic Medicine*, *65*(10), 611-621.

Tay, S. W., Ryan, P., & Ryan, C. A. (2016). Systems 1 and 2 thinking processes and cognitive reflection testing in medical students. *Canadian Medical Education Journal*, *7*(2), e97-e103. https://pubmed.ncbi.nlm.nih.gov/28344696

Wortman, P. M. (1972). Medical diagnosis: An information-processing approach. *Computers and Biomedical Research*, *5*(4), 315-328. https://doi.org/https://doi.org/10.1016/0010-4809(72)90065-1

Zimmerman, B. J. (2008). Investigating Self-Regulation and Motivation: Historical Background, Methodological Developments, and Future Prospects. *American Educational Research Journal*, *45*(1), 166-183. https://doi.org/10.3102/0002831207312909

Zwaan, L., & Hautz, W. E. (2019). Bridging the gap between uncertainty, confidence and diagnostic accuracy: calibration is key. *BMJ Quality & Safety*, *28*(5), 352-355.

# APPENDICES

## ENGLISH SUMMARY

This dissertation describes four studies aimed at improving the teaching of clinical reasoning in medical education. These studies answer two main questions. The studies described in Chapters 2-4 investigated the first main question. They tested whether participants can learn a specific procedure for *deliberate reflection*, so that they can apply it when diagnosing new cases. The study described in Chapter 5 investigated the second main question. It was tested whether participants could learn to better estimate how well they had diagnosed a case. The studies were conducted among students and physicians in general-practice training.

**Chapter 1** describes the theoretical background for the studies. *Deliberate reflection* is a procedure in which physicians are asked to systematically go through a number of steps in order to arrive at a medical diagnosis. First, they are asked to read a written case and come up with an initial diagnosis. Then, they are asked to list all the features from the case that (1) speak in favour of this diagnosis, (2) speak against this diagnosis, and (3) that are absent in this case, but which you would expect with the diagnosis at hand. Finally, they are asked to come up with an alternative diagnosis and to go through the steps for this diagnosis as well. In previous studies, these steps have been effective for diagnosing difficult cases and correcting diagnostic errors, for example, when these errors are the result of a misleading thought pattern (i.e., cognitive bias). For deliberate reflection to be effective not only in studies but also in medical practice, physicians must learn the procedure and then apply it autonomously when diagnosing new cases without being asked to do so by a researcher. Chapter 1 discusses several ways that may help to learn deliberate reflection. In addition, it is important that physicians recognize when they have made a wrong diagnosis, and when further reflection would be helpful. This relationship between the correctness of the diagnosis and the confidence a clinician has in the diagnosis is called *diagnostic calibration*. Chapter 1 also discusses a possible way to improve the diagnostic calibration of physicians in training.

**Chapter 2** describes a study conducted among physicians in general-practice training, which consisted of a learning phase and a test phase. In this study, we tested whether we could teach them deliberate reflection during the learning phase, so that they would apply it when diagnosing new cases in the test phase. In the learning phase, participants were assigned to one of three conditions: (1) control condition, (2) example-based learning condition, and (3) learning-by-doing condition. In the control condition, they had to diagnose cases without being taught deliberate reflection. In the example-based learning condition, they studied how an expert applied deliberate reflection to diagnose cases. In the learning-by-doing condition, participants first viewed examples of deliberate reflection to learn

how the procedure works and were then asked to apply the steps of deliberate reflection themselves when diagnosing cases.

The test phase was the same for all participants. Participants had to diagnose new medical cases and in some cases they were also asked to write down how they arrived at the diagnosis. We analysed whether their diagnoses in the test phase were correct and whether we could find elements of deliberate reflection in their reasoning. However, we found no differences between the conditions on any of these outcome measures. The interventions thus seemed ineffective for learning and applying deliberate reflection.

The study described in **Chapter 3** is a follow-up to the study in Chapter 2, as this study also tests whether physicians in general-practice training can learn deliberate reflection and then apply it autonomously. However, this study tests a different intervention to learn deliberate reflection and a different way to measure whether the participants use deliberate reflection than the study in Chapter 2. In the learning phase, participants were assigned to one of two conditions: (1) the control condition, or (2) the learning-by-teaching condition. In the control condition, they had to diagnose cases without being taught deliberate reflection. In the learning-by-teaching condition, they first studied examples of deliberate reflection and then were asked to record videos in which they explain to a fictitious peer what deliberate reflection is. An analysis of these videos showed that the participants had learned deliberate reflection. The test phase was the same for all participants. They were asked to diagnose new cases while thinking aloud. In this study, we again analysed whether their diagnoses in the test phase were correct and whether we could find elements of deliberate reflection in their reasoning. However, we found no differences between the conditions on any of these outcome measures.

This study shows that the intervention was effective for teaching deliberate reflection in the learning phase, but the participants did not apply the procedure when diagnosing new cases in the test phase. One of the possible explanations was that physicians in training to become general practitioners already have a lot of experience in diagnosing patients and therefore do not easily adapt their way of reasoning. Less experienced participants, such as medical students, may be more likely to adopt and apply deliberate reflection. Another possible explanation was that participants felt that the cases were easy and therefore it was not necessary to use deliberate reflection to find the correct diagnosis.

**Chapter 4** describes a study conducted among medical students, which consisted of a learning phase and a test phase. In this study, we tested whether we could teach participants the deliberate reflection procedure during the learning phase, so that they would apply it during the test phase when diagnosing new cases. We also tested whether students

would only use deliberate reflection if they expected a case to be difficult to diagnose. The learning phase followed the same procedure as the learning phase in the study described in Chapter 3; Students were assigned to (1) the control condition without deliberate reflection, or to (2) the learning-by-teaching condition, in which they explained deliberate reflection to a fictitious fellow student. The test phase was the same for all participants. Students were asked to diagnose new cases and then write down what they could remember about a case (recall task). After seeing half of the cases, they were told that the upcoming cases were difficult cases. In doing so, we wanted to give students the feeling that it would be useful to apply deliberate reflection, but the difficulty of the cases did not actually change.

By analysing what characteristics of the case the students were able to recall, we wanted to see what they had focused on while diagnosing the case. Some of the features in the case were related to one possible diagnosis but not to another possible diagnosis. The results showed that those students who had learned deliberate reflection also focused on the diagnosis that they had not given themselves. This may indicate that they used (some) steps of deliberate reflection, rather than just focusing on their own diagnosis. Contrary to our expectations, they also applied it when the cases were not described as difficult, which means that students did not need to be stimulated to apply deliberate reflection.

Finally, the study described in **Chapter 5** focuses on the second main questions of this thesis. This study tested the hypothesis that feedback on diagnostic performance (indicating whether someone diagnosed a case correctly or incorrectly) would improve physicians' diagnostic calibration, meaning they would be better in estimating how accurate their diagnosis was. In this study, residents from general-practice training were assigned to one of two conditions: (1) the feedback condition or (2) the no-feedback condition. Everyone diagnosed 12 written cases in one session. They gave a diagnosis and then rated how confident they were in their diagnosis. Subsequently, the participants in the feedback condition were shown the correct diagnosis for the case and were asked to compare this diagnosis with the diagnosis they had given themselves. Participants in the no-feedback condition did not get the correct answer, but went straight to the next case.

To analyse the results, we calculated the diagnostic calibration by calculating to what extent the correctness of their diagnosis matched the confidence in their diagnosis. The results showed that feedback did not help improve calibration. Further analyses showed that participants in the feedback condition were even less confident when they made a correct diagnosis than participants in the no-feedback condition. Therefore, it seemed that the feedback made them more uncertain, but did not help them distinguish between correct and incorrect diagnoses.

In **Chapter 6**, the findings of the four studies are discussed and integrated. The first two studies found no effect of learning deliberate reflection on the participants' reasoning when diagnosing new cases. In the third study, however, we did find an effect. A possible explanation for this is that the studies used different ways to measure the reasoning process. Another explanation is that medical students adopt deliberate reflection more quickly than physicians in the general-practice training. To improve diagnostic calibration, i.e., to better estimate whether a diagnosis was correct or incorrect, it does not seem effective to give physicians feedback on their previous diagnostic performance.

## NEDERLANDSE SAMENVATTING (DUTCH SUMMARY)

Dit proefschrift beschrijft vier studies die erop gericht zijn het onderwijs van klinisch redeneren in medische opleidingen te verbeteren. Deze studies beantwoorden twee vraagstukken. Het eerste vraagstuk wordt onderzocht in de studies omschreven in hoofdstukken 2-4. Hier werd getest of proefpersonen *deliberate reflection* (een procedure voor weloverwogen reflectie) kunnen leren, zodat ze die kunnen toepassen bij het diagnosticeren van nieuwe casussen. Het tweede vraagstuk wordt onderzocht in de studie omschreven in hoofdstuk 5. Hier werd getest of proefpersonen konden leren om beter in te schatten hoe goed ze een casus hebben gediagnosticeerd. De studies zijn uitgevoerd onder studenten en onder artsen in opleiding tot specialist (AIOS) in de huisartsopleiding.

**Hoofdstuk 1** beschrijft de theoretische achtergrond voor de studies. Met *deliberate reflection* wordt een procedure van bewuste, weloverwogen reflectie bedoeld, waarbij artsen gevraagd worden om een aantal stappen door te lopen om tot een medische diagnose te komen. Eerst worden ze gevraagd om een geschreven casus te lezen en een eerste diagnose te bedenken. Daarna worden ze gevraagd om een lijst te maken met alle kenmerken uit de casus die (1) voor deze diagnose spreken, (2) tegen deze diagnose spreken, en (3) die bij de patiënt afwezig zijn, maar die je zou verwachten bij deze diagnose. Tot slot worden ze gevraagd om een alternatieve diagnose te bedenken en ook voor deze diagnose de stappen te doorlopen. In eerdere studies hebben deze stappen kunnen helpen om moeilijke casussen te diagnosticeren en om diagnosefouten te verbeteren, bijvoorbeeld als deze fouten het resultaat zijn van een verkeerd gedachtenpatroon (d.w.z. cognitieve bias). Om deze voordelen van *deliberate reflection* niet alleen in studies maar ook in de medische praktijk te kunnen bereiken, moeten artsen de procedure aanleren en vervolgens zelfstandig toepassen bij het diagnosticeren van nieuwe casussen zonder dat een onderzoeker ze vraagt om dit te doen. In hoofdstuk 1 worden verschillende manieren besproken die zouden kunnen helpen om *deliberate reflection* aan te leren. Daarnaast is het belangrijk dat artsen herkennen wanneer ze een verkeerde diagnose hebben gemaakt, en wanneer verdere reflectie zou helpen. Deze relatie tussen de correctheid van de diagnose en het vertrouwen dat een arts in de diagnose heeft, noemen we diagnostische kalibratie. Hooftstuk 1 bespreekt ook een manier om de diagnostische kalibratie van artsen in opleiding te verbeteren.

**Hoofdstuk 2** beschrijft een studie uitgevoerd onder AIOS in de huisartsopleiding, die bestond uit een leerfase en een testfase. In deze studie we hebben getoetst of we hun *deliberate reflection* konden aanleren tijdens de leerfase, zodat ze deze zouden toepassen bij het diagnosticeren van nieuwe casussen in de testfase. In de leerfase werden de deelnemers verdeeld in één van drie condities: (1) controle conditie, (2) leren door voorbeelden conditie, en (3) leren door doen conditie. In de controle conditie moesten zij casussen

diagnosticeren, zonder dat aan hen *deliberate reflection* werd geleerd. In de leren door voorbeelden conditie (*example-based learning*) gingen ze bestuderen hoe een expert *deliberate reflection* toepaste om casussen te diagnosticeren. In de leren door doen conditie (*learning-by-doing*) bekeken de deelnemers eerst voorbeelden van *deliberate reflection* om te leren hoe de procedure werkt en werden ze vervolgens gevraagd om zelf de stappen *van deliberate reflection* toe te passen bij de diagnosticeren van casussen.

De testfase was voor alle deelnemers gelijk. Daarin moesten zij nieuwe casussen diagnosticeren en werden zij bij een deel ook gevraagd om achteraf op te schrijven hoe ze tot de diagnose waren gekomen. Wij hebben geanalyseerd of hun diagnoses in de testfase correct waren en of we elementen van de *deliberate reflection* konden terugvinden in hun redeneringen. Echter, op geen van deze uitkomstmaten bleken de condities te verschillen. De interventies leken dus niet effectief voor het leren en later toepassen van *deliberate reflection*.

De studie beschreven in **hoofdstuk 3** is een vervolg op de studie in hoofdstuk 2 gezien ook deze studie toetst of AIOS in de huisartsopleiding *deliberate reflection* kunnen leren en vervolgens zelfstandig toepassen. Deze studie toetst echter een andere interventie om *deliberate refelction* te leren en een andere manier om te meten of AIOS *deliberate reflection* toepassen dan de studie in hoofdstuk 2. In de leerfase werden de deelnemers verdeeld in één van twee condities: (1) de controleconditie, of (2) de leren door te onderwijzen conditie. In de controle conditie moesten zij casussen diagnosticeren, zonder dat hun *deliberate reflection* werd geleerd. In de leren door te onderwijzen conditie (*learning-by-teaching*) werd hun *deliberate reflection* door middel van voorbeelden geleerd en werden ze vervolgens gevraagd om video's op te nemen waarin ze aan een fictieve mede-AIOS uitleggen wat *deliberate reflection* is. Een analyse van deze video's liet zien dat de deelnemers *deliberate reflection* hadden geleerd. De testfase was voor alle deelnemers gelijk. Daarin werden ze gevraagd om nieuwe casussen te diagnosticeren terwijl ze hardop dachten. Wij hebben ook in deze studie geanalyseerd of hun diagnoses in de testfase correct waren en of we elementen van *deliberate reflection* konden terugvinden in hun redeneringen. Echter, op geen van deze uitkomstmaten bleken de condities te verschillen.

Terwijl we nu wisten dat de interventie effectief was voor het aanleren van *deliberate reflection* in de leerfase, pasten de deelnemers de procedure niet toe bij het diagnosticeren van nieuwe casussen in de testfase. Een van de mogelijke verklaringen was dat AIOS al veel ervaring met het diagnosticeren van patiënten hebben en daarom hun manier van redeneren niet zo makkelijk aanpassen. Minder ervaren deelnemers, zoals geneeskundestudenten, zouden *deliberate reflection* mogelijk sneller aannemen en toepassen. Een andere mogelijke verklaring was dat de deelnemers het gevoel hadden dat de casussen makkelijk

waren en dat het daarom niet nodig was om *deliberate reflection* toe te passen om de juiste diagnose te vinden.

**Hoofdstuk 4** beschrijft een studie uitgevoerd onder geneeskundestudenten, die bestond uit een leerfase en een testfase. Ook in deze studie hebben we getoetst of we tijdens de leerfase *deliberate reflection* konden leren aan de deelnemers, zodat ze deze in de testfase zouden toepassen bij het diagnosticeren van nieuwe casussen. Verder hebben we getoetst of studenten *deliberate reflection* alleen zouden toepassen, als ze verwachten dat een casus moeilijk te diagnosticeren zou zijn. De leerfase volgde dezelfde procedure als de leerfase in de studie beschreven in hoofdstuk 3; Studenten werden ingedeeld in (1) de controle conditie zonder *deliberate reflection*, of in (2) de leren door te onderwijzen conditie, waarin ze *deliberate reflection* gingen uitleggen aan een fictieve medestudent (*learning-by-teaching*). De testfase was voor alle deelnemers gelijk. Daarin werden de studenten gevraagd om nieuwe casussen te diagnosticeren en daarna op te schrijven wat ze zich nog konden herinneren van een casus (herinneringstaak). Nadat ze de helft van de casussen hadden gezien, werd hun verteld dat de komende casussen moeilijke casussen waren. Op deze manier wilden we de studenten het gevoel geven dat het zinvol was om de geleerde *deliberate reflection* procedure ook toe te passen, maar de moeilijkheid van de casussen veranderde niet daadwerkelijk.

Door te analyseren welke kenmerken van de casus de studenten later konden herinneren, wilden we kijken waarop ze zich tijdens het diagnosticeren van de casus hadden gefocust. Sommige van de kenmerken in de casus waren gerelateerd aan één bepaalde mogelijke diagnose maar niet aan een andere mogelijke diagnose. De resultaten toonden aan dat de studenten die *deliberate reflection* hadden geleerd, zich ook hadden gefocust op de diagnose die zij zelf niet hadden gegeven. Dit kan erop wijzen dat ze (sommige) stappen van *deliberate reflection* gebruikten, in plaats van zich alleen op hun eigen diagnose te concentreren. Tegen onze verwachting in pasten ze het ook toe wanneer de casussen niet als moeilijk werden beschreven, wat betekent dat studenten niet eerst gestimuleerd moesten worden om weloverwogen te reflecteren.

Tenslotte richt de studie beschreven in **hoofdstuk 5** zich op het tweede doel van dit proefschrift. Deze studie testte de hypothese dat feedback op de diagnostische prestaties (aangeven of iemand een casus goed of fout heeft gediagnosticeerd) de *diagnostische kalibratie* van artsen zou verbeteren, wat betekend dat ze beter kunnen inschatten hoe goed hun diagnose was. In deze studie werden AIOS uit de huisartsopleiding ingedeeld in één van twee condities: (1) de feedbackconditie of (2) de geen-feedbackconditie. Iedereen diagnosticeerde in één sessie 12 geschreven casussen. Ze stelden een diagnose en beoordeelden vervolgens hoeveel vertrouwen ze in hun diagnose hadden. Daarna kregen de

AIOS in de feedbackconditie de juiste diagnose voor de casus te zien en werden ze gevraagd deze diagnose te vergelijken met de diagnose die ze zelf hadden gegeven. Deelnemers in de geen-feedbackconditie kregen niet het juiste antwoord, maar gingen gelijk door naar de volgende casus.

Om de resultaten te analyseren, hebben we de diagnostische kalibratie berekend door te kijken in hoeverre de correctheid van hun diagnose paste bij het vertrouwen in hun diagnose. De resultaten toonden aan dat feedback niet hielp om de kalibratie te verbeteren. Verdere analyses lieten zien dat deelnemers in de feedbackconditie zelfs minder vertrouwen hadden wanneer ze een juiste diagnose hadden gesteld dan deelnemers in de geen-feedbackconditie. Het leek dus dat de feedback ze onzekerder maakte, maar hen niet hielp tussen correcte en incorrecte diagnoses te onderscheiden.

In **hoofdstuk 6** worden de bevindingen van de vier studies besproken en geïntegreerd. De eerste twee studies vonden geen effect van het leren van *deliberate reflection* op de manier van redeneren tijdens het diagnosticeren van nieuwe casussen. In de derde studie werd er wel een effect gevonden. Een mogelijke verklaring hiervoor is dat er in de studies verschillende manieren werden gebruikt om het redeneerproces te meten. Een andere verklaring is dat geneeskundestudenten *deliberate reflection* sneller aannemen dan AIOS van de huisartsopleiding. Voor het verbeteren van de diagnostische kalibratie, dus om artsen beter te laten inschatten of hun diagnose correct of incorrect was, lijkt het niet efficiënt om ze feedback te geven over hun eerdere diagnostische prestatie.

## CURRICULUM VITAE

Josepha Kuhn was born in Magdeburg, Germany on February 3, 1990. After finishing secondary education at the Gymnasium in Winsen Luhe, she moved to Groningen, The Netherlands, to study psychology. In 2013, she obtained her Bachelor of Science in Psychology from Rijksuniversiteit Groningen. After this, she moved to Rotterdam to start with the master program Brain and Cognition at Erasmus University Rotterdam. One year later, she started a second master Health Psychology at the University Leiden as well as a research internship at De Academische Werkplaats – De Nieuwe Kans where she helped with research among multi-problem youth in Rotterdam. In 2015, she obtained her Master's degree from Erasmus University Rotterdam and in 2016 her Master's degree from the University Leiden. In 2016 she started her PhD research at the institute of Medical Education Research (iMERR) and the General Practice Department of Erasmus Medical Centre. Her research focussed on ways to improve the education in clinical reasoning for medical students and physicians in general-practice training, and resulted in several national and international presentations and publications.

# PHD PORTFOLIO

| PhD Training and Activities | Year | Workload in ECs |
|---|---|---|
| **Courses and Workshops** | | |
| Clinical epidemiology | 2016 | 0.54 |
| Photoshop and Illustrator CS6 for PhD-students and other researchers | 2017 | 0.3 |
| Indesign CS6 for Phd-students and other researchers | 2017 | 0.15 |
| Bayesian Statistics and JASP | 2017 | 0.3 |
| Basic Course on 'R' | 2017 | 1.8 |
| Research Management | 2018 | 0.6 |
| Mindfulness - introduction course | 2018 | 1.2 |
| Research integrity | 2018 | 0.3 |
| Biomedical writing course | 2019 | 3 |
| Mindfulness - compassie training | 2019 | 0.9 |
| Endnote | 2019 | 0.13 |
| Learning how to learn | 2020 | 0.3 |
| Kunstgeschiedenis door Museumdirecteuren | 2021 | 0.64 |
| Klassiekers lezen met schrijvers | 2021 | 0.32 |
| | | |
| **Attending Conferences, Symposia, Meetings etc**. | | |
| iMERR lab meeting | 2016 – 2022 | 2 |
| DEM conference, Europe | 2016 | 0.3 |
| Projectleidersbijeenkomst HGOG | 2016 | 0.3 |
| NVMO conference | 2016 | 0.6 |
| Pub group | 2017 – 2019 | 0.5 |
| Klinisch redeneren symposium (HAG) | 2017 | 0.15 |
| NVMO promovendidag | 2017 | 0.3 |
| Open Science: the National Plan and you (van Science in Transition) | 2017 | 0.3 |
| JURE + EARLI conference | 2017 | 2.1 |
| DPECS Graduate Research Day | 2017 | 0.3 |
| NVMO conference | 2017 | 0.6 |
| Projectleidersbijeenkomst HGOG | 2017 | 0.3 |
| NHG wetenschapsdag | 2018 | 0.3 |
| iMERR Graduate Research Day | 2018 | 0.3 |
| DEM conference, Europe | 2018 | 0.6 |
| NVMO conference | 2018 | 0.6 |
| NVMO Promovendidag | 2019 | 0.3 |
| RIME conference | 2019 | 0.6 |

| PhD Training and Activities | Year | Workload in ECs |
|---|---|---|
| Symposium: Understanding Diagnostic Error | 2019 | 0.15 |
| DEM conference, USA | 2019 | 1.2 |
| NVMO Promovendidag | 2022 | 0.3 |
| | | |
| **Presentations** | | |
| iMERR lab meeting x 4 presentations | 2016 – 2022 | 2 |
| HAG Stafdag | 2016 | 0.5 |
| HAG kwartaallunch | 2016 | 0.5 |
| Projectleidersbijeenkomst HGOG | 2016 | 0.5 |
| Symposium at EARLI conference | 2017 | 0.5 |
| Lunch presentation for residents | 2017 | 0.5 |
| DPECS Graduate Research Day | 2017 | 0.5 |
| NVMO conference | 2017 | 0.5 |
| Projectleidersbijeenkomst HGOG | 2017 | 0.5 |
| NHG wetenschapsdag | 2018 | 0.5 |
| iMERR Graduate Research Day | 2018 | 0.5 |
| DEM conference, Europe | 2018 | 0.5 |
| NVMO conference | 2018 | 0.5 |
| RIME conference | 2019 | 0.5 |
| Symposium: Understanding Diagnostic Error | 2019 | 0.5 |
| DEM conference, USA x2 presentations | 2019 | 1 |
| | | |
| **Other activities** | | |
| Supervise bachelor theses | 2021 | 1.93 |
| Reviewer for *Diagnosis* | 2021 | 0.3 |