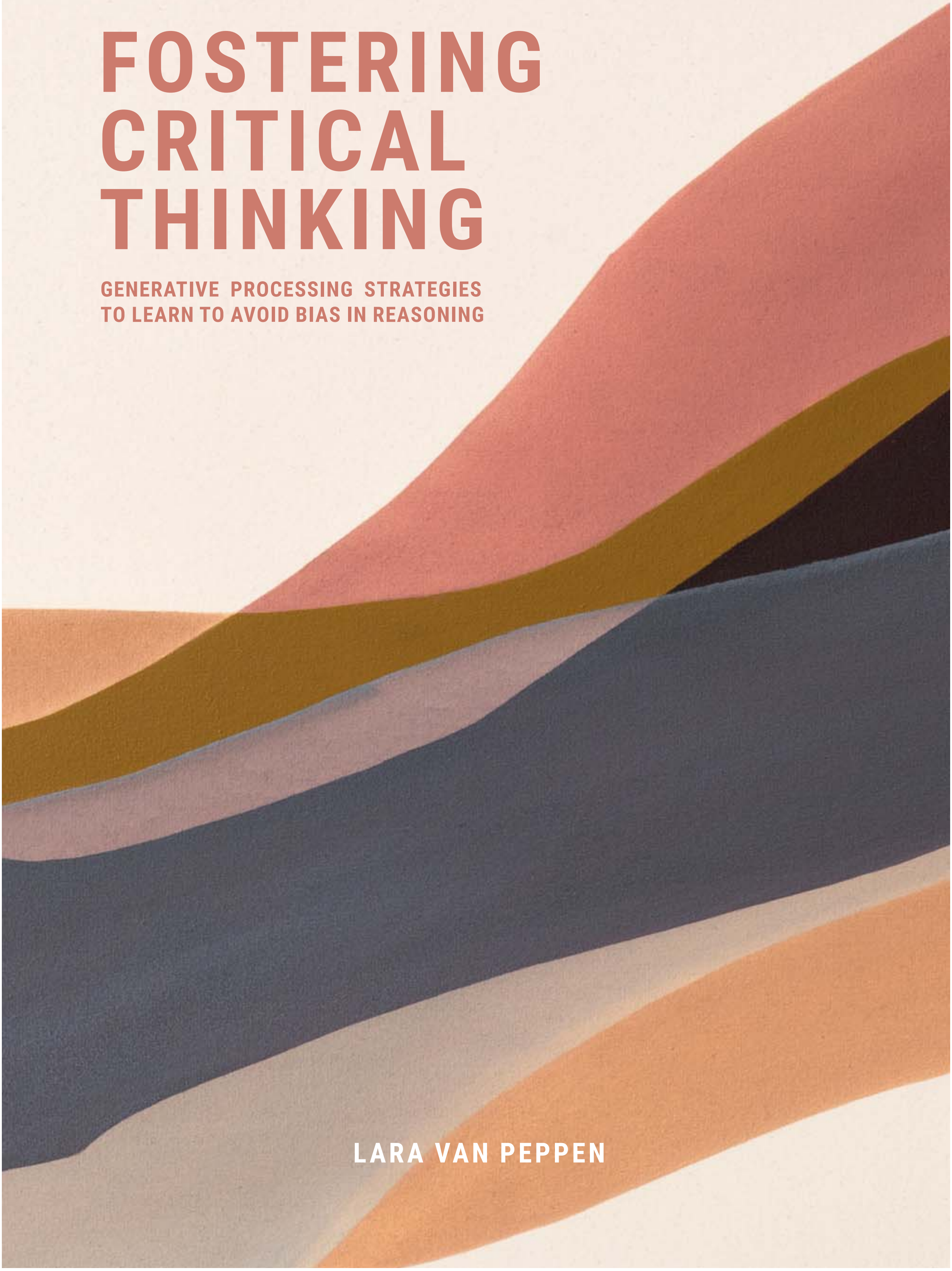
FOSTERING CRITICAL THINKING

GENERATIVE PROCESSING STRATEGIES TO LEARN TO AVOID BIAS IN REASONING



LARA VAN PEPPEN

FOSTERING CRITICAL THINKING

GENERATIVE PROCESSING STRATEGIES

TO LEARN TO AVOID BIAS IN REASONING

Colophon

The research reported in this dissertation was conducted in the context of the Interuniversity Center for Educational Sciences (ICO). It was funded by The Netherlands Organisation for Scientific Research (NWO project number 409-15-203) and co-financed by Avans University of Applied Sciences.

Cover design	Lara van Peppen, cover illustration *Above Land* (2020) by Lauren Mycroft
Lay-out	Lara van Peppen
Print	Ridderprint, BV, the Netherlands
ISBN	978-94-6416-112-0

Fostering Critical Thinking: Generative processing strategies to learn to avoid bias in reasoning

Bevorderen van kritisch denken: generatieve verwerkingsstrategieën om systematische redeneerfouten te leren vermijden

Proefschrift

ter verkrijging van de graad van doctor aan de
Erasmus Universiteit Rotterdam
op gezag van de rector magnificus

Prof.dr. R.C.M.E. Engels

en volgens het besluit van het College voor Promoties.

De openbare verdediging zal plaatsvinden op
vrijdag 25 september 2020 om 13:30 uur door

Lara Meike van Peppen
geboren te Delft

Erasmus University Rotterdam

Promotiecommissie

Promotor:
Prof.dr. T.A.J.M. van Gog

Overige leden:
Prof.dr. G.W.C. Paas
Prof.dr. P.W. van den Broek
Dr. K. Dijkstra

Copromotoren:
Dr. P.P.J.L. Verkoeijen
Dr. A.E.G. Heijltjes

Contents

Chapter 1

General introduction

Introduction

Every day, we make many decisions that are based on previous experiences and existing knowledge. This happens almost automatically as we rely on a number of heuristics (i.e., mental shortcuts) that ease reasoning processes (Tversky & Kahneman, 1974). Heuristic reasoning is typically useful, especially in routine situations. But it can also produce systematic errors (i.e., biases; this concept will be discussed in more detail later). Let us consider the following example:

> If someone conducts scientific research, s/he works at a university.
> Lara worked at the Erasmus University Rotterdam.
> Therefore, Lara conducted scientific research.

Because of its believability, most people will intuitively judge the conclusion as valid (cf. belief bias: Evans et al., 1983; Markovits & Nantel, 1989; Newstead et al., 1992), but the if-then rule does not state that someone working at a university conducts scientific research. We do know that Lara worked at a university, but we cannot deduce whether she conducted scientific research. Lara might have performed research procedures without scientific purposes, for instance, or she might have performed educational activities or support services. The conclusion would not necessarily follow from the premises and is, therefore, invalid. This syllogistic reasoning task requires replacement of the heuristic response with a response based on formal logic. Although in this example, the negative consequences are limited, heuristic reasoning can also produce biases with far-reaching consequences. To illustrate, a forensic expert who misjudges fingerprint evidence because it verifies his or her preexisting beliefs concerning the likelihood of the guilt of a defendant, displays the so-called confirmation bias, which can result in a misidentification and a wrongful conviction (e.g., the Madrid bomber case; Kassin et al., 2013). Fortunately, we are not doomed to reach wrong conclusions and to make incorrect decisions as in this example. Our primary tool for making better decisions is *critical thinking* (henceforth, in this dissertation, abbreviated as CT).

The importance of CT was already stressed by Socrates over 2,500 years ago and received renewed interest in the beginning of the 20th century. In 1910, John Dewey described the importance of critique and stated that *everyone* should engage in CT. Due to the expanding and changing demands that today's society places on people, the importance of being able to think critically has only increased (Pellegrino & Hilton, 2012). Because CT is essential for succeeding in future careers and to be efficacious citizens, helping students to become critically-thinking professionals is a central aim of higher education (Butler & Halpern, 2020; Davies, 2013; DeAngelo et al., 2009; Elen et al., 2019;

Facione, 1990; Halpern, 2014; Halpern & Butler, 2019; Van Gelder, 2005; Verburgh, 2013).

Consequently, many international (Ananiadou & Claro, 2009; OECD, 2018; Vincent-Lancrin et al., 2019) and national (i.e., Dutch: HBO-raad, 2009; OCW, 2019; Onderwijsraad, 2014a, 2014b, 2017, 2018; Vereniging Hogescholen, 2015) higher education policy documents include objectives to enhance students' CT-skills. To illustrate, around the start of this project, Avans Hogeschool, a Dutch University of Applied Sciences[1], had set explicit CT-aims in the documents detailing the educational ambitions (Avans Hogeschool, 2014a, 2014b) such as "every graduate is curious, shows a critical attitude, and is analytical. Therefore, we are committed to developing student's reflective and critical thinking capacity" (Avans Hogeschool, 2014b, p. 5)[2]. Several large-scale longitudinal studies, however, were quite pessimistic that this laudable goal would be realized merely by following a higher education degree program. These studies revealed that far too many higher education graduates lack the knowledge, beliefs, skills, and strategies required to think critically after four years of college (Arum & Roksa, 2011; Flores et al., 2012; Pascarella et al., 2011; although a more recent meta-analytic study reached the more positive conclusion that students' do improve their CT-skills over college years: Huber & Kuncel, 2016).

Hence, there is a growing body of literature on effective strategies for teaching CT in general (e.g., Abrami et al., 2008, 2014; Angeli & Valanides, 2009; Niu et al., 2013; Tiruneh et al., 2014, 2016) and avoiding reasoning biases in particular (Heijltjes et al., 2014a, 2014b, 2015; Van Brussel et al., 2020). It is well established, for instance, that bringing about learning of CT-skills is conditional upon provision of explicit CT-instructions and practice problems (e.g., Abrami et al., 2008, Heijltjes et al., 2014b). Yet there are still many open questions about optimal instructional designs to further enhance CT, and especially to establish transfer of CT-skills. Transfer refers to the ability to apply acquired knowledge and skills to novel situations (e.g., Barnett & Ceci, 2002; Perkins & Salomon, 1992). It is crucial that students think critically, especially in situations that have not been encountered before and where biases can have serious consequences (e.g., in complex professional environments in which the majority of higher education graduates are employed, such as medicine: Elia et al., 2016; Mamede et al., 2010; Law: Kassin et al., 2013; Koehler et al., 2002; Rachlinski, 2004). Therefore, the overall aim of this dissertation was to acquire more knowledge on effective strategies for fostering both learning and transfer of CT-skills in higher education, focusing specifically on avoiding bias in reasoning and drawing from findings from educational

[1] The Dutch education system distinguishes between research-oriented higher education (i.e., offered by research universities) and profession-oriented higher education (i.e., offered by universities of applied sciences).

[2] Surprisingly, CT is not (yet) explicitly mentioned in the latest ambition plan of Avans Hogeschool (2019).

and cognitive psychology. A brief overview of the history and theories of CT and biased reasoning and current research on teaching CT will serve as a preamble.

What is critical thinking?

CT finds its basis in the thoughts of Socrates, Plato, Aristotle, and other Greek philosophers. The term itself originated from this ancient Greek tradition as well; the word critical derives from the Greek words ‘kritikos’ (i.e., to judge/discern) and ‘kriterion’ (i.e., standards). Etymologically, then, CT implies making judgments based on standards. Hundreds of thinkers from different disciplines have subsequently made contributions to the idea of critical thought. John Dewey is considered the progenitor of the modern CT tradition. He described reflective thinking – his homologue to CT – to include “active, persistent, and careful consideration of any belief or supposed form of knowledge in the light of the grounds that support it, and the further conclusion to which it tends” (1910, p. 7). A variety of definitions has been suggested since then. Edward Glaser (1941), for example, expanded Dewey’s definition to recognize the role of having certain thinking skills, but also of being disposed to use these skills. He was the first to describe CT as a composite of attitudes, knowledge, and skills. Robert Ennis (1962) took Dewey's definition and transformed it into a more general simplified definition that could provide a basis for research. According to him, CT implies “reasonable reflective thinking focused on deciding what to believe or do”. The most accepted definition in the field of educational assessment and instruction, however, has been proposed by an expert Delphi Panel of the American Philosophical Association (APA). They agreed to characterize CT as:

> “purposeful, self-regulatory judgment that results in interpretation, analysis, evaluation, and inference, as well as explanations of the considerations on which that judgment is based… The ideal critical thinker is habitually inquisitive, well-informed, trustful of reason, open-minded, flexible, fair-mined in evaluation, honest in facing personal biases, prudent in making judgments, willing to reconsider, clear about issues, orderly in complex matters, diligent in seeking relevant information, reasonable in the selection of criteria, focused in inquiry, and persistent in seeking results which are as precise as the subject and the circumstances of inquiry permit.” (Facione, 1990, p. 2).

Despite the variety of definitions of CT and the multitude of components, there appears to be agreement that one key aspect of CT is the ability to avoid bias in reasoning and decision-making (Baron, 2008; Duron et al., 2006; West et al., 2008), which we will refer to as unbiased reasoning from hereon. This is the aspect of CT on which the research presented in this dissertation is focused. Bias is said to occur when a reasoning process results in a systematic deviation from ideal normative standards derived from laws of logic and probability (Stanovich et al., 2016; Tversky & Kahneman, 1974). Up to now, a substantial amount of literature has focused on the variety of heuristics and biases that exists. The so-called 'heuristics and biases' approach has generated influential research on CT and is central to this dissertation.

Heuristics and biases

The basic idea of the heuristics and biases approach, launched by Daniel Kahneman and Amos Tversky in the early 1970s, is that people rely on a variety of simple heuristics for judgment under uncertainty (Kahneman & Tversky, 1972, 1973; Tversky & Kahneman, 1974). As alluded to earlier, people resort to heuristics because these can help solve many different problems and make quick decisions, especially with rules and principles that have been practiced to automaticity (i.e., routine circumstances; Kahneman & Frederick, 2002; Kahneman & Klein, 2009; Shiffrin & Schneider, 1977; Stanovich, 2011). Heuristic reasoning allows us to not spend endless amounts of time and effort analyzing every information around us and is, therefore, very functional. For instance, when a medical emergency calls for action, an experienced clinician can use a recognizable pattern of cues to quickly make a diagnosis or size up a situation. However, the use of heuristics can also give rise to biases in reasoning and decision-making, as illustrated at the beginning of this chapter.

Kahneman and Tversky originally classified biases as associated with three such general-purpose heuristics (note that this is not the only classification of heuristics, however): representativeness, availability, and anchoring and adjustment (Tversky & Kahneman, 1974). The representativeness heuristic is characterized by the fact that people often evaluate the probability of an uncertain event by similarity with other events of the same type and causal/correlational beliefs (Chapman & Chapman, 1967; Jennings et al., 1982; Tversky & Kahneman, 1983). Specifically, representativeness concerns the degree of correspondence between an outcome and a model. To illustrate, Tversky and Kahneman (1983) asked undergraduates questions as "is it more probable that someone (selected by chance) has had a heart attack or that someone has had a heart attack and is over 55 years old?" Due to the natural assessment of a strong relation between heart failure and older age, thus high representativeness, the majority of graduates incorrectly perceived the conjunction of a heart attack and the age of 55 as more likely than a heart attack alone. Here, the use of the representativeness heuristic leads to neglect of conjunction rule ($P(A\&B) \leq P(B)$), known as the conjunction fallacy.

In case of the availability heuristic, people evaluate the probability of an event according to the ease with which examples come to mind (Tversky & Kahneman, 1973, 1983).

Events that are easy to retrieve from memory are regarded to be much more frequent and probable than they actually are. This can be the result of high exposure, as is the case, for instance, with terrorist attacks, airplane crashes, or natural disasters. But it can also be due to personal experiences/encounters. For example, if you are asked if it is more likely that the letter K appears in the first or third position of a word in English, you might estimate the first position as more probable. Just because it is much easier to recall words with the letter K in the first rather than the third position, however, the latter is actually more probable. In this case, the use of the heuristic results in availability bias (Tversky & Kahneman, 1973).

When people focus on an initial number or value (anchor) and then render a final estimate towards the anchor, they resort to the anchoring-and-adjustment heuristic (Lichtenstein & Slovic, 1971; Slovic, 1967; Tversky & Kahneman, 1974). For instance, when people were asked whether they would pay $25 (low anchor) or $200 (high anchor) to clean up lakes to protect fish populations and were then asked to estimate the amount the average person would contribute, they gave mean estimates of $14 and $36 with the low and high anchors, respectively. Here, the use of the heuristic leads to anchoring bias (Kahneman & Knetsch, 1993, Tversky & Kahneman, 1974).

Origins of biases

The occurrence of biases in thinking and reasoning can be explained by dual process models, which hold that there are two distinct cognitive systems that underlie thinking, reasoning, and decision-making: Type 1 and Type 2 processing, also referred to by some as System 1 and System 2 (Evans, 2003, 2008; Kahneman & Frederick, 2002; Stanovich, 1999, 2005, 2011). *Type 1 processing* is heuristic-based and operates automatically, autonomously, and rapidly, by means of parallel processing. As such, Type 1 processing is relatively effortless and does not place heavy demands on working memory. It has been shown that Type 1 processing is especially useful and functional during routine circumstances. Even a complex, but standard task can be completed with Type 1 processing (e.g., reading; or, for most Dutch people, cycling). However, in other (non-routine) situations, Type 1 processing might produce biased outcomes (Evans, 2003, 2008). Consider for example a clinician who has read information on a disease in the morning and later that day misdiagnoses a case of a patient who is presented with similar features (which triggered that diagnosis read earlier) but had in fact a different disease. That clinician makes use of the rapid and automatic Type 1 processes (i.e., availability heuristic, leading to availability bias; Schmidt et al., 2014). Thus, although the use of the availability heuristic may lead to efficient (i.e., fast and sound) decision-making in routine situations, it may also open the door to biases that could have been prevented by analytical and reflective reasoning, which is labelled as *Type 2 processing*.

Type 2 processing involves controlled processes that are relatively slow and largely sequential. One of the most crucial functions of Type 2 processing is to override Type 1 processing when this is to our benefit. To override Type 1 processing, one has to recognize the need for Type 2 processing and has to try to switch to this type of processing. This is only possible, however, when Type 1 processing can successfully be inhibited. Furthermore, this will only lead to a more favorable outcome when relevant mindware – consisting of both relevant procedural and conceptual knowledge – is available to provide better alternative responses (Aczel et al., 2015; Aron, 2008; Best et al., 2009; Stanovich, 2011; Zelazo, 2004). Biases occur when people use Type 1 processing when that is not appropriate, do not recognize the need for Type 2 processing, are not willing to switch to Type 2 processing or unable to sustain it (e.g., due to lack of sufficient cognitive capacity or time pressure), or miss the relevant mindware to come up with a better response. Consequently, in order to prevent biased reasoning, it is, first of all, necessary to stimulate people to switch to Type 2 processing. However, that may not be enough if they lack the necessary mindware, so in many cases, mindware has to be taught as well. In the next section, I will review what research has revealed with respect to effective ways of teaching CT in general, and then zoom in on effective methods for teaching students to avoid bias in reasoning.

Current research on teaching critical thinking

Previous research has established that CT-skills in general rarely evolve as a by-product of education; rather, they need to be explicitly taught (Abrami et al., 2008, 2014; Arum & Roksa, 2011; Beyer, 2008). However, there are different views of what the best way is to teach CT; the most well-known debate being whether CT should be taught in a general or content-specific manner (Abrami et al., 2014; Davies, 2013; Ennis, 1989; Moore, 2004). On the one hand, generalists (e.g., Ennis, 1989, 1992) ague that CT is a universal, general skill that can be applied to many contents and, as such, might be best learned separately from regular subject matter adjunct to the standard curriculum (Royalty, 1995; Stanovich & West, 1999). According to specifists (e.g., McPeck, 1990, 1992) on the other hand, CT cannot be separated from the subject matter to which it is applied and, therefore, should be taught in specific academic disciplines (Tsui, 2002). During the last years, this debate has faded away, since most researchers nowadays commonly agree that CT can be seen in terms of both general skills (e.g., sound argumentation, evaluating statistical information, and so on) and specific skills or knowledge used in the context of disciplines (e.g., Davies, 2013; Ikuenobe, 2001; Robinson, 2011; Smith, 2002; Tsui, 2002).

Indeed, it has been shown that the most effective teaching methods combine generic instruction, in which general CT-skills and dispositions are taught separately from subject matter, with the opportunity to integrate the general principles that were taught with domain-specific subject matter through infusion or immersion (i.e., mixed courses; for meta-analyses, see Abrami et al., 2008, 2014). In infusion methods, general CT principles are made explicit and students are encouraged to deal with specific subject matter in a critical way, while immersion methods invite students to reflect and make judgments on specific disciplinary issues without general CT principles made explicit (Ennis, 1989). Merely providing students with generic, infusion, or immersion courses, respectively, seemed less effective for fostering CT than mixed courses (Abrami et al., 2008, 2014). In the same vein, Tiruneh and colleagues (2014), found that both generic and mixed courses resulted in better CT outcomes than infusion and immersion courses.

Teaching for unbiased reasoning

A considerable number of studies on avoiding bias in reasoning has focused on strategies to mitigate specific biases (referred to as debiasing strategies; e.g., Aczel et al., 2015a; Catapano et al, 2019; Herzog & Hertwig, 2009; Kaufmann et al., 2010; Larrick, 2004; Lord et al., 1984). These studies, however, are not concerned with the implementation of these strategies in education. Some studies that did address teaching unbiased reasoning reflect the finding of studies concerned with teaching general CT-skills (Abrami et al., 2008, 2014): combining explicit CT-instruction with the opportunity to apply the principles that were taught on domain-relevant problems seems beneficial for learning of unbiased reasoning (Heijltjes et al., 2014a, 2014b, 2015). In these studies, students participated in a pretest-intervention-posttest design. The intervention consisted of either explicit, implicit, or no CT-instructions that were offered either with or without opportunity to practice in a domain context. Unbiased reasoning was operationalized as performance on classical heuristics-and-biases tasks (Tversky & Kahneman, 1974), in which an intuitively cued heuristic response conflicts normative models of CT as set by formal logic and probability theory.

Although these studies uncovered that a combination of explicit CT-instructions and task practice promotes learning of unbiased reasoning, they also consistently observed that this was not sufficient to establish transfer to novel problem types (and this also applies to CT-skills more generally, see for example, Halpern, 2014; Kenyon & Beaulac, 2014; Lai, 2011; Ritchhart & Perkins, 2005; Tiruneh et al., 2016). The process of transfer involves the application of acquired knowledge or skills to some new context or related materials (e.g., Barnett & Ceci, 2002; Cormier & Hagman, 2014; Druckman & Bjork, 1994; Haskell, 2001; McDaniel, 2007; Perkins & Salomon, 1992). In the educational psychology literature, transfer has been described as existing on a continuum from near to far, with lower degrees of similarity between the initial and transfer situation along the way (e.g., Perkins & Salomon, 1992). This lack of transfer is worrisome because it would be unfeasible to train students on each and every type of reasoning bias they will ever encounter (and this also applies to CT-skills more generally, see for example, Halpern, 2014; Kenyon & Beaulac, 2014; Lai, 2011; Ritchhart & Perkins, 2005). Surprisingly, though, it has not yet been investigated what kind of practice activities can promote (far) transfer.

The existing transfer literature suggests that, to establish transfer, instructional strategies should contribute to actively constructing meaning from to-be-learned information, by mentally organizing it in coherent knowledge structures and integrate these principles with one's prior knowledge (i.e., *generative processing*; Grabowski, 1996; Osborne & Wittrock, 1983; Wittrock, 1974, 1990, 1992, 2010). Generative processing can help learners acquire abstractions of the underlying principles behind a problem that are required for transfer of learned skills. If the potential transfer situation presents similar requirements and the learner recognizes them, they may select and apply the same or a somewhat adapted learned procedure to solve the novel problem (e.g., Gentner, 1983, 1989; Mayer & Wittrock, 1996; Reed, 1987; Vosniadou & Ortony, 1989). Indeed, strategies that encourage generative processing have been shown to foster knowledge acquisition and promote transfer of various other cognitive skills (e.g., Fiorella & Mayer, 2015, 2016). Ways to stimulate generative processing are, for instance, encouraging elaboration, questioning, or explanation during practice (e.g., Fiorella & Mayer, 2016; Renkl & Eitel, 2019), creating variability in practice (e.g., Barreiros et al., 2007; Moxley, 1979), stimulating comparison of correct problem solutions with erroneous ones (e.g., Durkin & Rittle-johnson, 2012; Loibl & Leuders, 2018, 2019), or having students repeatedly retrieve to-be-learned material from memory (Butler, 2010; Carpenter & Kelly, 2012; McDaniel et al., 2012, 2013; Rohrer et al., 2010).

Taken together, despite the value placed on teaching CT, it remains a disputed point how to do this more effectively. It has been established that bringing about learning of CT-skills is conditional upon provision of explicit CT-instructions and practice problems, but there are still many open questions about optimal practice activities to further enhance CT, in ways that transfer across tasks/contexts. To properly inform educational practice about optimally tailoring CT courses, further study is therefore required. The studies presented in this dissertation overall aim to gain more knowledge on fostering higher education students' learning and transfer of CT-skills – through instructional interventions that target generative processing – focusing specifically on avoiding bias in reasoning. This leads to the main research questions, which will be discussed in the next section.

Context and overview of this dissertation

This dissertation is one of the results of the broader NWO-funded research project "Investing in Thinking Pays Good Interest: Improving Critical Thinking Skills of Students and Teachers in Higher Professional Education". In this project, a consortium of researchers from Erasmus University Rotterdam and Utrecht University and educational policy advisors, teachers, and researchers from Avans University of Applied Sciences, aimed to generate knowledge on teaching CT that would be scientifically relevant as well as directly relevant for educational practice. The main objective of this project was to improve higher education students' CT-skills, by investigating how to equip teachers with the knowledge and skills needed to effectively teach unbiased reasoning (conducted by Eva Janssen, Utrecht University) and how to further enhance students' skills to avoid bias in reasoning, in such a way that these would also transfer across tasks/contexts.

The studies in **Chapters 2**, **3**, **4**, and **5** are concerned with the main question of whether instructional interventions that are known to foster generative processing and transfer of other cognitive skills, would further facilitate learning and transfer of CT-skills required for unbiased reasoning (i.e., above and beyond effects of instruction and practice). These interventions were administered after initial instruction, during the practice phase. In addition, the study in Chapter 6, experimentally examined what obstacle(s) prevent(s) successful transfer of these CT-skills. An important aspect of this dissertation is that all studies contained or consisted of an experiment conducted in a real educational setting and as part of an existing course (using educationally relevant materials) at a University of Applied Sciences, which increases ecological validity of the studies.

The classroom study presented in **Chapter 2** addressed the question of whether prompting students to self-explain during practice; that is, to generate explanations of a problem-solution to themselves (e.g., Bisra et al., 2018; Chi, 2000; Fiorella & Mayer, 2016) would be effective for fostering (transfer of) unbiased reasoning. Students were provided with instruction on the importance and features of CT, on the skills and attitudes needed to think critically, and on several heuristics-and-biases tasks. Subsequently, they performed practice activities on domain-relevant problems in the task categories they were given instructions on, either with or without self-explanation prompts. Students' performance on heuristics-and-biases tasks (both on instructed/practiced tasks, to assess learning, and on novel tasks that shared underling principles, to assess transfer), perceived mental effort investment, and time-on-test were measured on a pretest, immediate posttest, and two-week delayed posttest. Additionally, it was explored whether the quality of students' self-explanations was related to their performance.

In **Chapter 3**, two experiments (laboratory and classroom) tested whether creating variability during practice through interleaved practice (in which practice task categories vary from trial to trial, as opposed to blocked practice; e.g., Barreiros et al., 2007; Helsdingen et al., 2011; Rau et al., 2013) would be effective for fostering unbiased reasoning. While interleaved practice has been shown to enhance learning (e.g., Helsdingen et al., 2011a, 2011b) it is usually more cognitively demanding than blocked practice, and very high cognitive load may hinder learning (Paas et al., 2003a). Therefore, it was additionally examined whether learners would experience lower cognitive load and benefit more from interleaved practice, when using worked examples as opposed to practice problems (cf. Paas & Van Merriënboer, 1994). Worked examples have been shown to reduce ineffective cognitive load (compared to practice problems; Van Gog et al., 2019). After receiving explicit instruction on CT and specific heuristics-and-biases tasks, students either practiced in an interleaved schedule with worked examples, an interleaved schedule with problems, a blocked schedule with worked examples, or a blocked schedule with problems. Again, students' performance on several heuristics-and-biases tasks (both on instructed/practiced tasks and novel tasks), perceived mental effort investment, and time-on-test were measured on a pretest, immediate posttest, and two-week delayed posttest. Additionally, students' global judgements of learning and experienced cognitive load during practice were explored.

The classroom study reported in **Chapter 4** investigated whether comparing correct and erroneous examples (i.e., contrasting examples) would enhance unbiased reasoning more than studying correct examples only, studying erroneous examples only, and solving practice problems. Students were provided with the CT-instructions and practice on domain-relevant problems, under one of the four conditions. Their performance on heuristics-and-biases tasks (both on instructed/practiced tasks and novel tasks), mental effort investment, and time-on-test were measured on a pretest, immediate posttest, three-week delayed posttest, and nine-month delayed posttest. Furthermore, effects on perceived mental effort and time-on-task during practice were explored.

In **Chapter 5**, a classroom study is described that empirically investigated whether repeated retrieval practice over time (i.e., working on practice tasks in sessions that were weeks apart), would be beneficial for learning to reason in an unbiased manner and whether it can additionally facilitate transfer. Students were instructed on CT and avoiding belief-bias in syllogistic reasoning and practiced with syllogisms on domain-relevant problems, followed by feedback on their performance. Depending on assigned condition, they did not engage in extra practice, practiced a second time (week later), or practiced a second (week later) and third time (two weeks after second time).

Students' performance on heuristics-and-biases tasks (both on instructed/practiced syllogisms and novel tasks that shared similar features with syllogisms), mental effort investment, and time-on-test were measured on a pretest and immediate posttest. Additionally, explorative data on students' global judgements of learning, perceived mental effort during practice, time-on-task during practice, and time spent on worked-example feedback after correct and incorrect retrievals were collected.

Understanding the nuances of transfer is necessary to design courses to achieve it. So, it is crucial to gain insight into the obstacles to transfer of CT. Therefore, the study in **Chapter 6** focused exclusively on identifying whether unsuccessful transfer of CT-skills would be due to a failure to recognize that acquired knowledge is relevant in a new context, to recall that knowledge, or to apply that knowledge to the new context (i.e., the three-step model of transfer; Barnett & Ceci, 2012). In two experiments (classroom and laboratory), students received explicit instructions on CT and avoiding belief-bias in syllogistic reasoning and practiced with syllogisms on domain-relevant problems. Students' performance on heuristics-and-biases tasks (on syllogisms with different story contexts to assess learning, syllogisms in a different format to assess near transfer, and novel tasks that shared similar features with syllogisms to assess transfer) and time-on-test were measured on a pretest and immediate posttest. On the posttest transfer items, students received no support, received recognition support, were prompted to recall the acquired knowledge, or received recall support (cf. Butler et al., 2013, 2017). The effects of support for different steps in the process were compared to infer where difficulties arise for learners. Additionally, it was explored (within the free recall condition) whether students' ability to recall the acquired knowledge was related to their posttest performance on near and transfer items.

Finally, **Chapter 7** provides a summary and discussion of the main findings of Chapters 2 to 6. In addition, this chapter discusses the implications for future research on CT and for educational practice.

Chapter 2

Effects of self-explaining on learning and transfer of critical thinking skills

This chapter has been published as:
Van Peppen, L. M., Verkoeijen, P. P. J. L., Heijltjes, A. E. G., Janssen, E. M., Koopmans, D., & Van Gog, T. (2018). Effects of self-explaining on learning and transfer of critical thinking skills. *Frontiers in Educational Psychology, 3*, 100. https://doi.org/10.3389/feduc.2018.00100

Abstract

Critical thinking is considered to be an important competence for students and graduates of higher education. Yet, it is largely unclear which teaching methods are most effective in supporting the acquisition of critical thinking skills, especially regarding one important aspect of critical thinking: avoiding biased reasoning. The present study examined whether creating desirable difficulties in instruction by prompting students to generate explanations of a problem-solution to themselves (i.e. self-explaining) is effective for fostering learning and transfer of unbiased reasoning. Seventy-nine first-year students of a Dutch Applied University of Sciences were first instructed on two categories of 'heuristics-and-biases' tasks (syllogism and base-rate or Wason and conjunction). Thereafter, they practiced these either with (self-explaining condition) or without (no self-explaining condition) self-explanation prompts that asked them to motivate their answers. Performance was measured on a pretest, immediate posttest, and delayed (two weeks later) posttest on all four task categories, to examine effects on learning (performance on practiced tasks) and transfer (performance on non-practiced tasks). Participants' learning and transfer performance improved to a comparable degree from pretest to immediate posttest in both conditions, and this higher level of performance was retained on the delayed posttest. Surprisingly, self-explanation prompts had a negative effect on posttest performance on practiced tasks when those were Wason and conjunction tasks, and self-explaining had no effect on transfer performance. These findings suggest that the benefits of explicit instruction and practice on learning and transfer of unbiased reasoning cannot be enhanced by increasing the difficulty of the practice tasks through self-explaining.

Introduction

Fostering students' critical thinking (CT) skills is an important educational objective, as these skills are essential for effective communication, reasoning and problem-solving abilities, and participation in a democratic society (Billings & Roberts, 2014). Therefore, it is alarming that many higher education students find it hard to think critically; their level of CT is often too low (Flores et al., 2012) and CT-skills do not seem to improve over their college years (e.g., Arum & Roksa, 2011). As early as 1910, John Dewey described the importance of critique and stated that *everyone* needs to engage in CT. A variety of CT definitions has been suggested since then, the most accepted definition in the field of educational assessment and instruction of which has been proposed by an expert Delphi Panel of the American Philosophical Association (APA; Facione, 1990). They characterized CT as "purposeful, self-regulatory judgment that results in interpretation, analysis, evaluation, and inference, as well as explanation of the evidential, conceptual, methodological, criteriological, or contextual considerations on which that judgment is based" (Facione, 1990, p.2). Despite the variety of definitions of CT and the multitude of components, there appears to be agreement that one key aspect of CT is the ability to avoid biases in reasoning and decision-making (West et al., 2008), which we will refer to as unbiased reasoning from hereon. Bias is said to occur when a reasoning process results in a systematic deviation from a norm when choosing actions or estimating probabilities (Stanovich et al., 2016; Tversky & Kahneman, 1974). As biased reasoning can have serious consequences in situations in both daily life and the complex professional environments (e.g., economics, law, and medicine) in which the majority of higher education graduates end up working, it is essential to teach unbiased reasoning in higher education (e.g., Koehler et al., 2002; Rachlinski, 2004). However, it is still largely unclear how unbiased reasoning can be best taught, and especially how *transfer* can be fostered; that is, the ability to apply acquired knowledge and skills to new situations (e.g., Davies, 2013).

In line with findings of research on teaching CT in general (e.g., Abrami et al., 2014), previous research on unbiased reasoning has shown that providing students with explicit instructions and giving them the opportunity to practice what has been learned, improves performance on the learned tasks, but not transfer (e.g., Heijltjes et al., 2014b). This lack of transfer is a problem, as it is important that students can apply what has been learned to other situations. According to the *desirable difficulties* framework (e.g., Bjork, 1994; Bjork & Bjork, 2011; Fyfe & Rittle-Johnson, 2017; Soderstrom & Bjork, 2015), long-term performance and transfer can be enhanced by techniques that are effortful during learning and may seem to temporarily hold back performance gains. Conditions that support rapid improvement of performance (i.e. retrieval strength) often only support momentary performance gains and do not contribute to permanent changes needed for learning (Bjork & Bjork, 2011). To enhance long-term retention and transfer of learned skills, storage strength should be increased by effortful learning conditions that trigger deep processing (Yan et al., 2016). The active and deeper processing produced by encountering desirable difficulties can promote transfer to new situations (cf. germane load; Soderstrom & Bjork, 2015). If, however, the difficulties evoke learners to invest additional effort on processes that are not directly relevant for learning or the learners miss the relevant knowledge or skills to successfully deal with them, they become undesirable (McDaniel & Butler, 2010; Metcalfe, 2011).

Although conditions inducing the most immediate and observable signs of performance improvements are often preferred by both teachers and learners because they appear to be effective, it is important for teachers and students alike to search for conditions that confront students with desirable difficulties and thereby facilitate learning and transfer (Bjork et al., 2015). Such conditions include, for example, spacing learning sessions apart rather than massing them together (i.e., spacing effect), mixing practice-task categories rather than practicing one task category before the next (i.e., interleaving effect), and testing learning material rather than simply restudying it (i.e., testing effect; e.g., Weissgerber et al., 2018). Another desirable difficulty is the active generation of an answer, solution, or procedure rather than the mere passive reception of it (i.e., generation effect; for a review see Bertsch et al., 2007). Generative processing of learning materials requires learners to invest additional effort on the learning processes and to be actively involved in these processes, such as encoding and retrieval processes (Yan et al., 2016). Therefore, generative learning activities contribute to the connection and entrenchment of new information from the to-be-learned materials to existing knowledge. As a result, understanding of the materials is stimulated and is more likely to be recallable at a later time or in a different context (Bjork & Bjork, 2011; DeWinstanley & Bjork, 2004; Fiorella & Mayer, 2016; Slamecka & Graf, 1978).

One promising strategy to promote generative learning, and thus to create desirable difficulty in instruction, is *self-explaining* (e.g., Fiorella & Mayer, 2016). Self-explaining involves the generation of explanations of a problem-solution to oneself rather than simply answering tasks passively. Indeed, self-explaining has been shown to foster knowledge acquisition and to promote transfer in a variety of other domains (e.g., Dunlosky et al., 2013; Fiorella & Mayer, 2016; Lombrozo, 2006; Rittle-Johnson & Loehr, 2017; Wylie & Chi, 2014; for a review see Bisra et al., 2018), but the effectivity in CT-instruction is not yet clear. Self-explaining is assumed to lead to the construction of meaningful knowledge structures (i.e., mindware), by investing effort in identifying knowledge gaps or faulty mental models and connecting new information to prior knowledge (e.g., Atkinson et al., 2003; Chi, 2000; Fiorella & Mayer, 2016), and seems especially effective in domains guided by general underlying principles (Rittle-Johnson & Loehr, 2017). Moreover, self-explaining might stimulate students to stop and think about new problem-solving strategies (Siegler, 2002) with engagement in more analytical and reflective reasoning, labeled as Type 2 processing, as a result. This type of processing is required to avoid biases in reasoning and decision-making. Biases often result from relying on Type 1 processing to solve problems, which is a relatively effortless, automatic, and intuitive type of processing. Although Type 1 processing may lead to efficient decision-making in many routine situations, it may open the door to errors that could have been prevented by engaging in Type 2 processing (e.g., Evans, 2008; Stanovich, 2011). As such, self-explaining might contribute to decoupling prior beliefs from available evidence, which is an essential aspect of unbiased reasoning. It is important to bear in mind, however, that the benefit of self-explaining only applies when students are able to provide self-explanations of sufficient quality (Schworm & Renkl, 2007).

Several studies demonstrated that prompting self-explaining fostered learning and/or transfer of certain aspects of CT-skills, such as argumentation (e.g., Schworm & Renkl, 2007), complex judgments (e.g., Helsdingen et al., 2011b), or logical reasoning (e.g., Berry, 1983). Studies on the effect of self-explanation prompts on unbiased reasoning (Heijltjes et al., 2014a, 2014b, 2015), however, showed mixed findings. One study found an effect on transfer performance on an immediate posttest (Heijltjes et al., 2014a), but this effect was short-lived (i.e., not retained on a delayed posttest) and not replicated in other studies (Heijltjes et al., 2014b; Heijltjes et al., 2015). This lack of (prolonged) effects of self-explaining might have been due to the nature of the final tests, which were multiple-choice (MC) answers only. A study in which students had to motivate their MC-answers suggests that this might provide a better, more sensitive measure of the effects of self-explaining on transfer of unbiased reasoning (Hoogerheide et al., 2014). Therefore, the present study used MC-plus-motivation tests to investigate whether self-explaining is effective for fostering learning and transfer of unbiased reasoning.

Since it seems reasonable to assume, but is as yet unproven, that increasing the desirable difficulty of learning materials through self-explaining might foster learning and transfer of unbiased reasoning, the present study was conducted as part of an existing critical thinking course (i.e., classroom study) to examine the usefulness of this desirable difficulty in a real educational setting. We investigated the effects of self-explaining during practice with 'heuristics-and-biases tasks' (e.g., Tversky & Kahneman, 1974) on learning and transfer, as assessed by final test tasks which required students to motivate their MC-answers. Based on the literature reviewed above, we hypothesized that explicit CT-instructions combined with practice on domain-specific cases would be effective for learning: therefore, we expected performance gains on practiced tasks from pretest to posttest as measured by MC-answers (Hypothesis 1). The more interesting question, however, is whether self-explaining during practice would lead to higher performance gains on practiced (i.e., *learning*; Hypothesis 2a) and non-practiced tasks (i.e., *transfer*; Hypothesis 2b) than not being prompted to self-explain during practice. As outlined before, we expected that beneficial effects of self-explaining on performance outcomes are more likely to be detected when participants are required to motivate their answer to MC-items. We hypothesized that self-explaining during practice would lead to higher total posttest scores (i.e., MC-plus-motivation) on practiced (i.e., *learning*; Hypothesis 3a) and non-practiced tasks (i.e., *transfer*; Hypothesis 3b). We expected this pattern of results to persist on the delayed posttest.

Furthermore, we explored perceived mental effort investment in the test items to get more insight into the effects of self-explaining on learning (Question 4a) and transfer performance (Question 4b). On the one hand, it can be expected that the acquisition of knowledge of rules and strategies would lower the cognitive load imposed by the task, and therefore participants might have to invest less mental effort on the posttests than on the pretest (Paas et al., 2003a), especially after having engaged in self-explaining. On the other hand, as both our training-phase and the self-explanation prompts were designed to provoke Type 2 processing – which is more effortful than Type 1 processing (Evans, 2011) – participants might have been inclined to invest *more* effort on the posttests than on the pretest, especially on the non-practiced (i.e., transfer) tasks, on which participants had not acquired any knowledge during instruction. Finally, because the quality of self-explanations has been shown to be related to learning and transfer, we explored whether the quality of the self-explanations on the practice tasks correlated with the immediate and delayed posttest performance (Question 5).

Materials and methods

We created an Open Science Framework (OSF) page for this project, where all materials, a detailed description of the procedure, and the dataset of the experiment are provided (osf.io/85ce9).

Participants and design

Participants were all first-year 'Safety and Security Management' students of a Dutch University of Applied Sciences (N = 88). Five participants missed the second session and four participants failed to complete the experiment due to technical problems. Therefore, the final sample consisted of 79 students (M_{age} = 19.16, SD = 1.61; 44 males). Because this study took place in a real educational setting and was part of an existing course, our sample was limited to the total number of students in this cohort. In response to a reviewer, we added a power function of our analyses using the G*Power software (Faul et al., 2009). The power of our 3×2×2 mixed ANOVAs – under a fixed alpha level of .05, with a correlation between measures of .3, and with a sample size of 79 – is estimated at .36, .99, and > .99 for picking up a small, medium, and large interaction effect, respectively. Regarding our 2×2×2 mixed ANOVAs, the power is estimated at .32, .96, and > .99 for picking up a small, medium, and large interaction effect, respectively. The power of our study, thus, should be sufficient to pick up medium-sized effects, which is in line with the mean weighted medium effect size of self-explaining of previous studies as indicated in a recent meta-analysis (Bisra et al., 2018).

The experiment consisted of four phases: pretest, training-phase (CT-instructions plus practice), immediate posttest, and delayed posttest (see Table 1 for an overview). Participants were randomly assigned to one of two conditions: (1) self-explaining condition (CT-instructions and CT-practice with self-explanation prompts; $n = 39$) and (2) no self-explaining condition (CT-instructions and CT-practice without self-explanation prompts; $n = 40$). Of the four task categories tested in the pretest and posttests participants received instruction and practice on two task categories (one involving statistical and one involving logical reasoning, see section CT- skills tests). To ensure that any condition effects would not be due to specific characteristics of the instructed and practiced tasks, half of the participants in each condition got instruction and practice on the first logical and the first probabilistic reasoning task category (i.e., syllogism and base-rate), and the other half on the second logical and the second probabilistic reasoning task category (i.e., Wason and conjunction).

Materials

CT-skills tests

The pretest consisted of eight classic heuristics-and-biases tasks that reflected important aspects of CT across four categories (i.e., two of each category): 1) *Syllogistic reasoning* tasks, which examine the tendency to be influenced by the believability of a conclusion when evaluating the logical validity of arguments (adapted from Evans, 2003); 2) *Wason selection* tasks, that measure the tendency to verify rules rather than to falsify them (adapted from Stanovich, 2011); 3) *Base-rate* tasks, which measure the tendency to overrate individual-case evidence (e.g., from personal experience, a single case, or prior beliefs) and to underrate statistical information (adapted from Fong et al., 1986; Tversky & Kahneman, 1974); and 4) *Conjunction* tasks, that measure to what extent people neglect a fundamental rule in probability theory, that is, the conjunction rule ($P(A\&B) \leq P(B)$) which states that the probability of Event A and Event B both occurring must be lower than the probability of Event A or Event B occurring alone (adapted from Tversky & Kahneman, 1983). The syllogistic reasoning and Wason selection tasks involve logical reasoning (i.e., Wason selection tasks can be solved by applying modus ponens and modus tollens from syllogistic reasoning) and the base-rate and conjunction tasks involve statistical reasoning (i.e., both require knowledge of probability and data interpretation). The content of the surface features (cover stories) of all test items was adapted to the study domain of the participants. A multiple-choice format with four answer options was used, with only one correct answer, except for one base-rate task where two answers were correct.

The immediate and delayed posttests were parallel versions of the pretest (i.e., structurally equivalent tasks but with different surface features). During the posttests, participants were additionally asked to motivate their MC-answers ("Why is this answer correct? Explain in steps how you have come to this answer.") by typing their motivation in a text entry box below the MC-question. The posttest items on the practiced task categories served to assess differences in learning outcomes, whereas the posttest items on the non-practiced task categories served to assess transfer performance. The transfer task categories shared similar features with the learning categories, namely, one requiring knowledge and rules of logic (i.e., syllogisms rules) and one requiring knowledge and rules of statistics (i.e., probability and data interpretation).

Table 1. Overview of the study design

	Self-explaining		No self-explaining	
	A (n = 18)	B (n = 21)	C (n = 22)	D (n = 18)
Pretest	Syllogism, Wason, Base-rate, and Conjunction	Syllogism, Wason, Base-rate, and Conjunction	Syllogism, Wason, Base-rate, and Conjunction	Syllogism, Wason, Base-rate, and Conjunction
Training phase Instruction and practice (version)	Syllogism and Base-rate	Wason and Conjunction	Syllogism and Base-rate	Wason and Conjunction
Self-explanation prompts during practice (condition)	Yes	Yes	No	No
Immediate posttest	Syllogism, Wason, Base-rate, and Conjunction	Syllogism, Wason, Base-rate, and Conjunction	Syllogism, Wason, Base-rate, and Conjunction	Syllogism, Wason, Base-rate, and Conjunction
Delayed posttest	Syllogism, Wason, Base-rate, and Conjunction	Syllogism, Wason, Base-rate, and Conjunction	Syllogism, Wason, Base-rate, and Conjunction	Syllogism, Wason, Base-rate, and Conjunction

CT-instructions

The text-based CT-instructions consisted of a general instruction on deductive and inductive reasoning and explicit instructions on two of the four categories from the pretest, including two extensive worked examples (of the tasks seen in the pretest) of each category. Participants received the following hints stating that the principles used in these tasks can be applied at several reasoning tasks: “Remember that these reasoning schemes can be applied in several reasoning tasks” and “Remember that the correct calculation of probabilities is an important skill that can be applied in several reasoning tasks”.

CT-practice

The CT-practice phase consisted of a case (315 words text) – on a topic that participants might encounter in their working-life – and four practice problems, two of each of the two task categories that students were given instructions on. In the self-explanation condition, participants were exposed to a self-explanation prompt after each of these tasks in which they were asked to explain how the answer was obtained: “Why is this answer correct? Explain in steps how you have come to this answer”.

Mental effort

After each test item participants reported how much mental effort they invested in completing that item, on a 9-point rating scale ranging from (1) very, very low effort to (9) very, very high effort (Paas, 1992; Paas & Van Merriënboer, 1993).

Procedure

The study was run during the first two lessons of a CT-course in the Safety and Security Management study program of an institute of higher professional education and conducted in the classroom with an entire class of students present. Participants signed an informed consent form at the start of the experiment. All materials were delivered in a computer-based environment (Qualtrics platform) that was created for this experiment, except for the paper-based case during the CT-instructions. The Qualtrics program randomly assigned the participants to a condition/version. Participants could work at their own pace, were allowed to use scrap paper while solving the tasks, and time-on-task was logged during all phases.

The study consisted of two sessions. In session 1 (during the first lesson of the course, ca. 90 min.), participants first completed the pretest. Subsequently, they had to read the CT-instructions and the case, followed by the practice problems, which differed according to the assigned condition/version. At the end, participants completed the immediate posttest. After two weeks, session 2 (during the second lesson of the course, ca. 30 min.) was held in which participants completed the delayed posttest. Invested mental effort was rated after each test item on all CT-skills tests. Both the teacher and the experiment leader (first author of this paper) were present during all phases of the experiment.

Scoring

For selecting a correct MC-answer on the three CT-skills tests, 1 point was assigned, resulting in a maximum MC-score of four points on the learning (i.e., instructed/practiced task categories) items and four points on the transfer (i.e., task categories not instructed/practiced) items on each test. On the immediate and delayed posttest, participants were additionally asked to motivate their MC-answers. These motivations were scored based on a coding scheme that can be found on our OSF page. In addition to the MC-score (1 point), participants could earn a maximum of two points per question for the given motivation, resulting in a maximum total score (MC-plus-motivation) of three points per item. Because one syllogism task had to be removed from the tests due to an inconsistent variant in the delayed posttest (i.e., relatively easier form), participants who received instructions on the syllogistic reasoning and base-rate tasks, could attain a maximum total score of nine on the learning items and 12 on the transfer items on each posttest; and vice versa for the participants who received instructions on the Wason and conjunction tasks. For comparability, we computed percentage scores on the learning and transfer items instead of total scores. Two raters independently scored 25% of the immediate posttest. The intra-class correlation coefficient was .952 for the learning test items and .971 for the transfer test items. Because of these high inter-rater reliabilities, the remainder of the tests was scored by one rater.

The quality of participants' explanations was determined on the basis of the self-explanations given during the practice tasks with a maximum of two points per task (cf. posttest explanation-scoring procedure). As there were four practice tasks, the maximum self-explanation score was eight (ranging from 0 to 8). Two raters independently scored 25% of the tasks. Because the inter-rater reliability was high (intra-class correlation coefficient of .899), the remainder of the tasks was scored by one rater.

Results

For all analyses in this paper a *p*-value of .05 was used as a threshold for statistical significance. Partial eta-squared (η_p^2) is reported as a measure of effect size for the ANOVAs, for which .01 is considered small, .06 medium, and .14 large, and Cohen's *d* is reported for the post-hoc tests, with values of 0.20, 0.50, and 0.80 representing a small, medium, and large effect size respectively (Cohen, 1988).

Preliminary analyses confirmed that there were no significant differences between the conditions before the start of the experiment in educational background, $\chi^2(3) = 2.41$, $p = .493$, gender, $\chi^2(1) = 0.16$, $p = .900$, or performance, time-on-task, and mental effort on the pretest (all *F*s < 1, maximum $\eta_p^2 = .01$). An independent-samples t-test indicated, surprisingly, that there were no significant differences in time-on-task (in seconds) spent on practice of the instruction tasks between the self-explaining condition ($M = 409.25$, $SD = 273.45$) and the no self-explaining condition ($M = 404.89$, $SD = 267.13$), $t(77) = 0.07$, $p = .943$, $d = 0.02$.

Test performance

Data are provided in Table 2 and test statistics in Table 3. Regarding the version of the instruction, only main effects of Version or interactions of Version with other factors are reported. The remaining results are provided in Table 3.

Performance gains on MC-answers

To test hypotheses 1, 2a, and 2b, two 3×2×2 mixed ANOVAs were conducted with Test Moment (pretest, immediate posttest, and delayed posttest) as within-subjects factor and Self-explaining (self-explaining and no self-explaining) and Version (syllogism and base-rate: SB, and Wason selection and conjunction: WC) as between-subjects factors.

Test Moment significantly affected *learning* (i.e., performance on practiced tasks): performance was lower on the pretest ($M = 40.40$, $SD = 29.09$) than on the immediate posttest ($M = 78.06$, $SD = 26.22$), $p < .001$, $\eta_p^2 = .65$. Performance on the immediate posttest did not differ significantly from that on the delayed posttest ($M = 79.54$, $SD = 25.17$), $p = .611$, $\eta_p^2 < .01$. Note though, that there was an interaction between Test Moment and Version; participants who received the SB-version showed an immediate to delayed posttest performance gain ($M_{immediate} = 74.16$; $M_{delayed} = 78.28$), whereas the WC-version showed a slight performance drop ($M_{immediate} = 82.54$; $M_{delayed} = 81.45$); however, follow-up tests showed that the gain and drop were non-significant, $F(1, 38) = 13.12$, $p = .001$, $\eta_p^2 = .26$; $F(1, 37) = 0.07$, $p = .794$, $\eta_p^2 = .002$. There was no main effect of Self-explaining nor an interaction between Test Moment and Self-explaining, indicating that prompting self-explanations did not affect learning gains.

There was a main effect of Test Moment on test performance on *transfer* (i.e., non-practiced) items. Performance was lower on the pretest ($M = 36.71$, $SD = 27.07$) than on the immediate posttest ($M = 49.37$, $SD = 30.16$), $p < .001$, $\eta_p^2 = .17$, which in turn was lower than on the delayed posttest ($M = 58.02$, $SD = 29.07$), $p = .004$, $\eta_p^2 = .11$. There was a main effect of Version: receiving the WC-version resulted in higher transfer performance (*M* = 57.98, *SE* = 3.46) than the SB-version (*M* = 38.47, *SE* = 3.42), indicating that transfer from WC-tasks to SB-tasks was higher than from SB-tasks to WC-tasks. Moreover, there was an interaction between Test Moment and Version. Follow-up analyses showed an effect of Test Moment for both the SB-version, $F(2, 76) = 10.74$, $p < .001$, $\eta_p^2 = .22$, and the WC-version, $F(2, 74) = 16.58$, $p < .001$, $\eta_p^2 = .31$. The pretest to immediate posttest performance gain was only significant for the SB-version, $F(1, 38) = 16.32$, $p = .001$, $\eta_p^2 = .30$, whereas the immediate to delayed posttest performance gain was only significant for the WC-version, $F(1, 37) = 17.64$, $p < .001$, $\eta_p^2 = .32$. There was no main effect of Self-explaining nor a significant interaction between Test Moment and Self-explaining, indicating that prompting self-explanations did not affect transfer performance.

Effects of self-explaining on learning outcomes (MC-plus-motivation)

To test hypothesis 3a, we analyzed the data of the MC-plus-motivation scores on learning items using a 2×2×2 mixed ANOVA with Test Moment (immediate posttest and delayed posttest) as within-subjects factor and Self-explaining (self-explaining and no self-explaining) and Version (syllogism and base-rate: SB, and Wason selection and conjunction: WC) as between-subjects factors (see Table 2 and 3 for data and test statistics, respectively). Pretest scores were not included in this analysis because the pretest only consisted of MC-questions. There was no main effect of Test Moment. Self-explaining significantly affected performance on learning items. Surprisingly, performance was higher in the no self-explaining condition (*M* = 64.36, *SE* = 3.26), compared to the self-explaining condition (*M* = 54.87, *SE* = 3.30). Note though, that there was an interaction between Self-explaining and Version. The effect of self-explaining was only found for the WC-version, $F(1, 37) = 7.66$, $p = .009$, $\eta_p^2 = .17$; there was no main effect of self-explaining for the SB-version, $F(1, 38) = 0.01$, $p = .953$, $\eta_p^2 < .01$. We did not find an interaction between Test Moment and Self-explaining.

Effects of self-explaining on transfer performance (MC-plus-motivation)

To test hypothesis 3b, we analyzed the data of the MC-plus-motivation scores on the transfer items using a 2×2×2 mixed ANOVA with Test Moment (immediate posttest and delayed posttest) as within-subjects factor and Self-explaining (self-explaining and no self-explaining) and Version (syllogism and base-rate: SB, and Wason selection and conjunction: WC) as between-subjects factors (see Table 2 and 3 for data and test statistics, respectively). There were no main effects of Test Moment and Self-explaining nor an interaction between Test Moment and Self-explaining. Collectively, the results on the transfer items again suggest that transfer occurred to a comparable extent in the self-explaining condition and the no self-explaining condition. Note though, that there was a main effect of version of instruction. In line with the findings on the MC-scores data, performance was higher for the WC-version (*M* = 49.95, *SE* = 3.31) than the SB-version (*M* = 25.60, *SE* = 3.27), indicating that transfer was higher when instructed or practiced with the WC-tasks compared to the SB-tasks.

Mental effort investment

Again, data are provided in Table 2 and test statistics in Table 3. We exploratively analyzed the mental effort data (average mental effort invested per learning item) using two 3×2×2 mixed ANOVAs with Test Moment (pretest, immediate posttest, and delayed posttest) as within-subjects factor and Self-explaining (self-explaining and no self-explaining) and Version (syllogism and base-rate: SB, and Wason selection and conjunction: WC) as between-subjects factors (Question 4a and 4b). Regarding the version of the instruction, only main effects of Version or interactions of Version with other factors are reported. The remaining results are available in Table 3. One participant had more than two missing values and was removed from the analysis.

There were no main effects of Test Moment or Self-explaining on effort invested in *learning* items, nor an interaction between Test Moment and Self-explaining. Note tough, that there was a main effect of version of instruction. Less effort investment on learning items was reported for the WC-version (*M* = 3.65, *SE* = 0.17) than the SB-version (*M* = 4.52, *SE* = 0.17). Moreover, there was an interaction between Self-explaining and Version. The effect of self-explaining was only found for the WC-version, $F(1, 36) = 5.08$, $p = .030$, $\eta_p^2 = .12$; there was no main effect of self-explaining for the SB-version, $F(1, 38) = 1.26$, $p = .268$, $\eta_p^2 = .03$.

Regarding effort invested in *transfer* items, Mauchly's test indicated that the assumption of sphericity had been violated, $\chi 2$ (2) = 7.45, $p = .024$, and therefore Huynh-Feldt corrected tests are reported ($\varepsilon = .95$). Mental effort was affected by Test Moment. Invested mental effort was lower on the pretest (*M* = 3.98, *SE* = 0.14) compared to the immediate posttest (*M* = 4.63, *SE* = 0.15), $p < .001$, $\eta_p^2 = .21$, which did not differ from that on the delayed posttest (*M* = 4.38, *SE* = 0.17), $p = .160$, $\eta_p^2 = .03$. There was no main effect of Self- explaining nor an interaction between Test Moment and Self-explaining

Quality of self-explanations

Several authors have reported that self-explanations are only beneficial when the quality of the explanations is sufficient (e.g., Schworm & Renkl, 2007). To examine whether we could corroborate this finding, we conducted an exploratory analysis. Based on the quality of the self-explanations in the instruction tasks, we created three groups: (1) highest self-explanation scores (score ≥ 4; 25% of the total group), (2) scores between

2 and 3 (42% of the total group), and (3) lowest self-explanation scores (score ≤ 1; 33% of the total group). We examined whether the quality of the self-explanations was related to performance on the learning (practiced) items by conducting a mixed ANOVA (on participants in the self-explanation condition) with Test Moment (immediate posttest and delayed posttest) as within-subjects factor and Quality of Self-explanations (high, medium, and low) as between-subjects factor. There was no main effect of Test Moment, $F(1, 36) = 0.02$, $p = .881$, $\eta_p^2 < .01$, but there was a main effect of Quality of Self-explanations, $F(2, 36) = 8.79$, $p = .001$, $\eta_p^2 = .33$. The group with the lowest self-explanation scores performed lower on learning items ($M = 36.86$, $SE = 5.38$) than the group with the medium self-explanation scores ($M = 59.55$, $SE = 4.85$), $p < .001$. The group with the medium self-explanation scores did not differ from the group with the highest self-explanation scores ($M = 69.17$, $SE = 6.13$), $p = .226$. No interaction between Test Moment and Quality of Self-explanations was found, $F(2, 36) = 1.26$, $p = .297$, $\eta_p^2 = .06$.

A similar mixed ANOVA was conducted to explore whether the quality of the self-explanations was related to performance on the transfer (non-practiced) items. There was no main effect of Test Moment, $F(1, 36) = 2.73$, $p = .107$, $\eta_p^2 = .07$, no main effect of Quality of Self-explanations, $F(2, 36) = 0.01$, $p = .994$, $\eta_p^2 < .01$, nor an interaction between Test Moment and Quality of Self-explanations, $F(2, 36) = 0.61$, $p = .550$, $\eta_p^2 = .03$.

Table 2. Means (*SD*) of Test performance (multiple-choice % score), Test performance (multiple-choice plus motivation % score), and Mental effort (1–9) per Condition and Version

		Self-explaining			No self-explaining		
		A	B	Total	C	D	Total
Learning items							
Test performance (MC)	Pretest	55.56 (28.01)	26.19 (23.02)	39.74 (29.15)	57.58 (25.58)	20.83 (19.65)	41.04 (29.38)
	Immediate	74.07 (31.43)	76.19 (26.78)	75.21 (28.64)	74.24 (25.05)	88.89 (19.60)	80.83 (23.66)
	Delayed posttest	77.78 (22.87)	72.62 (31.53)	75.00 (27.64)	78.78 (24.22)	90.28 (17.44)	83.96 (21.96)
Test performance (MC-plus-motivation)	Immediate	58.64 (23.43)	51.59 (25.77)	54.84 (24.65)	60.61 (20.35)	68.06 (22.55)	63.96 (21.42)
	Delayed posttest	62.04 (24.53)	47.22 (26.26)	54.06 (23.39)	59.34 (18.50)	69.44 (20.01)	63.89 (19.61)
Mental effort	Pretest	4.30 (1.13)	4.20 (1.27)	4.25 (1.19)	4.52 (1.14)	3.61 (1.10)	4.11 (1.20)
	Immediate	3.98 (1.27)	3.68 (1.29)	3.82 (1.27)	4.80 (1.54)	3.18 (1.00)	4.07 (1.54)
	Delayed posttest	3.91 (1.24)	4.19 (1.78)	4.05 (1.53)	4.02 (1.42)	3.58 (1.49)	3.58 (1.49)
Transfer items							
Test performance (MC)	Pretest	25.00 (24.25)	44.44 (30.43)	35.47 (29.10)	29.55 (23.95)	48.15 (23.49)	37.92 (25.24)
	Immediate	40.28 (28.62)	46.03 (32.45)	43.38 (30.48)	48.86 (27.25)	62.96 (30.01)	55.21 (29.03)
	Delayed posttest	41.67 (30.92)	66.67 (25.82)	55.13 (30.63)	45.45 (25.16)	79.63 (16.72)	60.83 (27.55)
Test performance (MC-plus-motivation)	Immediate	22.69 (21.92)	37.30 (27.22)	30.56 (25.68)	26.89 (21.51)	55.86 (23.45)	39.93 (26.49)
	Delayed posttest	25.93 (23.55)	45.50 (24.95)	36.47 (25.95)	26.89 (19.23)	61.11 (18.86)	42.29 (25.52)
Mental effort	Pretest	4.01 (1.28)	4.20 (1.27)	4.11 (1.26)	4.05 (1.28)	3.61 (1.10)	3.85 (1.21)
	Immediate	4.42 (1.40)	4.47 (1.09)	4.44 (1.23)	5.17 (1.41)	4.37 (1.22)	4.81 (1.37)
	Delayed posttest	4.53 (1.41)	4.67 (1.69)	4.60 (1.48)	4.45 (1.48)	3.81 (1.56)	4.17 (1.53)

Note. Instructional conditions: Version A and C = instructed on and practiced with syllogistic reasoning and base-rate tasks; Version B and D = instructed on and practiced with Wason and conjunction tasks

Table 3. Results mixed ANOVAs

	Performance (MC-only)			Performance (MC+motivation)			Mental Effort		
	F-test (df)	*p**	η_p^2	F-test (df)	*p**	η_p^2	F-test (df)	*p**	η_p^2
Learning									
Test Moment	98.13 (2,150)	<.001*	.57	0.01 (1,75)	.925	<.01	2.67 (2,148)	.073	.04
Self-explaining	1.21 (1,75)	.274	.02	4.19 (1,75)	.044*	.05	0.57 (1,74)	.455	.01
Version	2.82 (1,75)	.097	.04	0.05 (1,75)	.817	<.01	6.46 (1,74)	.013*	.08
Test Moment × Self-explaining	1.57 (2,150)	.212	.02	0.02 (1,75)	.903	<.001	2.20 (2,148)	.115	.03
Test Moment × Version	24.53 (2,150)	<.001*	.25	0.32 (1,75)	.571	<.01	2.03 (2,148)	.135	.03
Self-explaining × Version	0.72 (1,75)	.397	.01	4.52 (1,75)	.037*	.06	5.61 (1,74)	.020*	.07
Test Moment × Self-explaining × Version	1.99 (2,150)	.141	.03	1.34 (1,75)	.250	.02	0.36 (1,148)	.697	.01
Transfer									
Test Moment	23.36 (2,150)	<001*	.24	3.63 (1,75)	.061	.05	7.03 (1.94,148.00)	.001*	.09
Self-explaining	3.00 (1,75)	.088	.04	1.97 (1,75)	.164*	.03	0.33 (1,74)	.565	<.01
Version	16.09 (1,75)	<.001*	.18	27.36 (1,75)	<.001*	.27	1.10 (1,74)	.297	.02
Test Moment × Self-explaining	0.93 (2,15)	.399	.01	0.50 (1,75)	.482	.01	2.90 (1.94,148.00)	.060	.04
Test Moment × Version	4.81 (2,150)	.009*	.06	1.36 (1,75)	.248	.02	0.27 (1.94,148.00)	.760	<.01
Self-explaining × Version	0.33 (1,75)	.569	<.01	2.43 (1,75)	.124	.03	2.48 (1,74)	.119	.03
Test Moment × Self-explaining × Version	0.38 (2,150)	.682	.01	0.00 (1,75)	.974	.000	0.06 (1.94,148.00)	.939	.001

*$p < .05$

Discussion

Previous research has shown that creating desirable difficulty in instruction by having learners generate explanations of a problem-solution to themselves (i.e., self-explaining) rather than simply answering tasks passively, is effective to foster learning and transfer in several domains (Fiorella & Mayer, 2016). Regarding unbiased reasoning, Heijltjes and colleagues (2014a) demonstrated that self-explaining during practice had a positive effect on transfer of unbiased reasoning, but this effect was short-lived and not replicated in other studies (Heijltjes et al., 2014b, 2015). However, these findings were based on MC-answers only, and there are indications that effects of self-explaining on transfer may be detected when more sensitive MC-plus-motivation tests are used (Hoogerheide et al., 2014). With the present experiment, we aimed to find out whether instruction followed by self-explaining during practice with heuristics-and-biases tasks would be effective for learning and transfer, using final tests that required participants to motivate their MC-answers.

Consistent with earlier research, our results corroborate the idea that explicit CT-instruction combined with practice is beneficial for learning to avoid biased reasoning (Hypothesis 1), as we found pretest to immediate posttest gains on practiced tasks, remaining stable on the delayed posttest after two weeks, as measured by performance on the MC-only questions. This is in line with the notion that the acquisition of relevant mindware contributes to an adequate use of Type 2 processing which can prevent biased reasoning (Stanovich et al., 2008). Contrary to earlier findings (e.g., Heijltjes et al., 2014a), our experiment seemed to provide some evidence that these instructions and practice tasks may also enhance transfer. However, this only applied to participants who practiced with the syllogism and base-rate version. For participants who received the other version, transfer performance gains were reached at a later stage. As such, this may mean that either transfer was easier from syllogism and base-rate to Wason and conjunction or, given that this pattern is not consistent across analyses, that our findings may reflect non-systematic variance. Another reason why caution is warranted in interpreting this finding is that the maximum scores differed per version, which, even though we used percentage scores, might be an issue for comparability.

As for our main question, we did not find any indications that prompting self-explanations to increase the difficulty of the practice tasks had a differential effect – compared to the control condition – on learning (Hypothesis 2a) or transfer (Hypothesis 2b) performance gains. Nor did the analyses of the MC-plus-motivation data show a benefit of prompting self-explanations during practice for learning (Hypothesis 3a) or transfer (Hypothesis 3b). Surprisingly, our findings even suggest that self-explaining during practice may actually be less beneficial for learning: participants who received self-explanation prompts benefitted less from the instructions than those who were not prompted; however, this was only the case for one of the versions, so again, this finding needs to be interpreted with caution.

The findings of the present study are contrary to previous studies that demonstrated that self-explaining is effective for establishing both learning and transfer in a variety of domains (for a review see Fiorella & Mayer, 2016), but they are in line with the studies on unbiased reasoning (which assessed performance only by means of MC-answers) that demonstrated no positive effects (Heijltjes et al., 2014b, 2015) or only a short-lived effect of self-explaining on transfer (Heijltjes et al., 2014a). We did find that learners who gave lower quality self-explanations also performed worse on the learning items on the test (Question 5), which seems to corroborate the idea that a higher quality of self-explanations is related to higher performance (Schworm & Renkl, 2007), but it is possible that this finding reflects a priori knowledge or ability difference rather than an effect of the quality of self-explanations on performance. Thus, this study (with a more extensive performance measure) contributes to a small body of evidence that self-explanation prompts seem to have little or no benefit for acquiring unbiased reasoning skills.

One possible reason for the lack of a self-explanation effect could be the fact that the learners did not receive feedback on their self-explanations given in the practice phase. Providing feedback after students' self-explanations could have contributed to consolidating correct explanations and correcting or elaborating incorrect or incomplete explanations (e.g., Hattie & Timperley, 2007), which is of great importance in the domain of unbiased reasoning—arguably even more so than in other learning domains.

Another possibility might be that the nature of the tasks moderates effects of self-explaining. Contrary to previous studies, transfer on the tasks in the present study relies not only on deep understanding of the domain-specific knowledge involved in the task, but also on the ability to inhibit Type 1 processing and to switch to Type 2 processing. Possibly, prompting students to self-explain did not provoke the 'stop and think' reaction that was needed for transfer above and beyond what the instructions already accomplished. Our findings regarding effort investment support this idea (i.e., higher effort investment on transfer items on the posttests compared to the pretest in both conditions), suggesting that our training-phase provoked Type 2 processing, but there was no (additional) effect of the self-explanation prompts on effort investment.

A strength of the present study worth mentioning, is that – contrary to previous studies (e.g., Chi et al., 1994) – both conditions spent equal time on the practice tasks. Hence, it could be hypothesized that the beneficial effects of self-explaining in these studies are not direct but caused by mediation: generating explanations usually requires more time and spending more time on subject matter increases performance. According to this hypothesis, the effect of self-explaining should disappear when time-on-task is equated between the conditions. Indeed, Matthews and Rittle-Johnson (2009) observed that solving tasks with self-explanations and solving more tasks without explanations in the same amount of time, resulted in equal final test performance. However, there are mixed results within the few studies that equated time-on-task, with some studies finding beneficial effects of self-explaining, while others did not (e.g., De Bruin et al., 2007; De Koning et al., 2011; McEldoorn et al., 2013; Matthews & Rittle-Johnson, 2009) and most other studies on self-explaining did not (fully) report time-on-task (see Bisra et al., 2018). Thus, there is a definite need for more research that examines the interplay between self-explanation, time-on-task, and final test performance.

Another possibility why we did not find effects of self-explaining on learning of unbiased reasoning skills, however, is that our study was conducted as part of an existing course and the learning materials were part of the exam. Because of that, students of the control condition may have imposed desirable difficulties on themselves, for instance by covertly trying to come up with explanations for the questions. It seems likely that students would be more willing to invest effort when their performance on the learning materials actually matters (intrinsically or extrinsically) for them, which is often the case in field experiments conducted in real classrooms where the learning materials are related to the students' study domain. Therefore, it is possible that effects of desirable difficulties such as self-explaining found in the psychological laboratory – where students participate to earn required research credits and the learning materials are not part of their study program and sometimes even unrelated to their study domain – might not readily transfer to classroom studies. This would explain why previous studies, which are mostly laboratory studies, demonstrated effects of self-explaining and why these effects were mostly absent and in one case only short-lived in the classroom studies on unbiased reasoning (e.g., Heijltjes et al., 2014a, 2014b, 2015). Moreover, this finding suggests a theoretical implication, namely that beneficial effects of creating desirable difficulty in instruction might become smaller when the willingness to invest increases and vice versa.

Future work might investigate why self-explanation prompts as used in the present study seem to have no additional effect after instruction and practice and whether strategies to improve students' quality of self-explanations would have beneficial effects on learning, and especially, transfer performance. Enhancing the quality of the self-explanations could be accomplished by, for example, providing students with a self-explanation training in advance or by providing prompts that include some instructional assistance (cf. Berthold et al., 2009). Moreover, future research could investigate via classroom studies whether other desirable difficulties would be more beneficial for establishing learning and transfer of unbiased reasoning. In contrast to prompting self-explanations, other desirable difficulties such as creating task variability during practice and spacing of learning sessions apart, may result in beneficial effects since students of the control conditions cannot impose these desirable difficulties themselves (e.g., Weissgerber et al., 2018).

To conclude, based on the findings from the present study in combination with prior studies, prompting to self-explain during practice does not seem to be promising to enhance unbiased reasoning skills. This suggests that the nature of the task may be a boundary condition for effects of self-explaining on learning and transfer. Moreover, this study raises the question whether effects of self-explaining depend on the setting of the study, and thus contribute to knowledge about the usefulness of desirable difficulties in real educational settings. Considerably more research is needed to investigate how unbiased reasoning should be taught and especially how transfer can be fostered. This is important, because biased reasoning can have huge negative consequences in situations in both daily life and complex professional environments.

Chapter 7

Summary and general discussion

As outlined in the introduction, it is essential that higher education students are trained to become critically-thinking professionals. Critical thinking (CT) is crucial for succeeding in future careers and, moreover, is an important life skill (Davies, 2013; Facione, 1990; Halpern, 2014; Van Gelder, 2005). More specifically, students should be able to avoid biases in their reasoning and decision-making, even in situations that have not been encountered before, because especially in (professional) situations, reasoning biases can have serious consequences. However, it would be unfeasible to train students on each and every type of reasoning bias they will ever encounter. The challenge for educational practitioners, therefore, is to design instruction/practice so that students acquire the necessary resources to enhance CT in such a way that it would also transfer across tasks and contexts.

The overarching purpose of the research presented in this dissertation was to acquire more knowledge on how higher education students' learning and transfer of CT-skills can be fostered, focusing specifically on one important aspect of CT: the ability to avoid bias in reasoning. The studies in Chapters 2 to 5 examined whether instructional interventions that are known to foster generative processing and transfer of other cognitive skills, would further enhance learning and transfer of CT-skills required for unbiased reasoning (i.e., above and beyond effects of instruction and practice). These interventions were administered after initial instruction, during the practice phase. Through generative processing, learners actively construct meaning from to-be-learned information, by mentally organizing it in coherent knowledge structures and integrating these principles with existing knowledge (Grabowski, 1996; Osborne & Wittrock, 1983; Wittrock, 1974, 1990, 1992, 2010), which is required for transfer of learned skills. In addition, the study presented in Chapter 6 experimentally examined what obstacle(s) prevent(s) successful transfer of these CT-skills. In this final chapter, the main findings of the studies are presented and positioned in the broader literature first. Subsequently, the implications for research and educational practice are discussed, along with potential directions for future research.

Summary of main findings

The main question addressed in this dissertation, was to investigate how (i.e., which instructional strategies could be used) to further enhance learning of unbiased reasoning and to establish transfer to novel problem types. Instructional strategies that encourage generative processing seem to hold a considerable promise with respect to deep learning and transfer of other cognitive skills (e.g., Fiorella & Mayer, 2016; Wittrock, 2010), but their effects on the acquisition of CT-skills had hardly been investigated. The generative processing strategies investigated in the studies presented in this dissertation were: prompting students to self-explain during practice (Chapter 2), creating variability in practice (Chapter 3), stimulating comparison of correct problem solutions with erroneous ones (Chapter 4), and having students repeatedly retrieve to-be-learned material from memory (Chapter 5). In all of these studies, students participated in a pretest-intervention-posttest design. During the intervention, students were provided with instruction on the importance and features of CT, on the skills and attitudes needed to think critically, and on specific heuristics-and-biases tasks. Subsequently, they performed practice activities on domain-relevant problems in the task category/categories they were given instructions on, either with or without the respective generative processing strategies. Unbiased reasoning has been operationalized as performance on classic heuristics-and-biases tasks (Tversky & Kahneman, 1974), in which an intuitively cued heuristic response conflicts normative models of CT as set by formal logic and probability theory. Students' performance (both on task categories that were part of the practice phase to assess learning and novel task categories that share underlying principles to assess transfer) and perceived mental effort were measured on a pretest and posttest. Additionally, Chapters 2 to 4 included delayed posttests.

The classroom study presented in **Chapter 2** addressed the question of whether prompting students to self-explain during practice; that is, to generate explanations of a problem-solution to themselves (e.g., Bisra et al., 2018; Chi, 2000; Fiorella & Mayer, 2016) would be effective for fostering (transfer of) unbiased reasoning. Students were provided with instruction on the importance and features of CT, on the skills and attitudes needed to think critically, and on several heuristics-and-biases tasks. Subsequently, they performed practice activities on domain-relevant problems in the task categories they were given instructions on, either with or without self-explanation prompts. Results revealed that learning outcomes improved after instruction/practice (i.e., from pretest to posttest) and remained stable after a two-week delay. In contrast to previous findings in a variety of domains (for a review see Bisra et al., 2018), however, prompting self-explanations had no differential effect – compared to the control condition that did not receive prompts – on learning gains or transfer performance. Remarkably, mental effort investment did not differ across conditions. That raises the possibility that students in the control condition had also engaged in generative processing, for instance by covertly trying to come up with explanations for the questions. Additionally, it was explored whether the quality of students' self-explanations was related to their performance. Results indicated that this was the case: learners who gave lower quality self-explanations also performed worse on the learning (but not on transfer) items on the test, which seems to corroborate the idea that a higher quality of self-explanations is related to higher performance (Schworm & Renkl, 2007). It is possible, however, that this finding reflects a priori knowledge or ability difference rather than an effect of the quality of self-explanations on performance.

In **Chapter 3**, two experiments (laboratory and classroom) tested whether creating variability during practice through *interleaved practice* (in which practice task categories vary from trial to trial, as opposed to blocked practice; e.g., Barreiros et al., 2007; Helsdingen et al., 2011; Rau et al., 2013) would be effective for fostering unbiased reasoning. While interleaved practice has been shown to enhance learning (e.g., Helsdingen et al., 2011a, 2011b) it is usually more cognitively demanding than blocked practice, and a very high cognitive load may hinder learning (Paas et al., 2003a). Therefore, it was additionally examined whether learners would experience lower cognitive load and benefit more from interleaved practice, when using worked examples as opposed to practice problems (cf. Paas & Van Merriënboer, 1994). Worked examples have been shown to reduce ineffective cognitive load (compared to practice problems; Van Gog et al., 2019). After receiving explicit instruction on CT and specific heuristics-and-biases tasks, students either practiced in an interleaved schedule with worked examples, an interleaved schedule with problems, a blocked schedule with worked examples, or a blocked schedule with problems. In both experiments, learning outcomes again improved after instruction/practice (i.e., from pretest to posttest). However, contrary to expectations and previous findings (e.g., Barreiros et al., 2007; Likourezos et al., 2019; Moxley, 1979), there were no indications that interleaved practice led to better learning or transfer than blocked practice, irrespective of task format. Interestingly, the laboratory experiment demonstrated a benefit of studying worked examples over solving problems on learning outcomes, reached with less effort during the tests (i.e., more effective and efficient: Hoffman & Schraw, 2010; Paas & Van Merriënboer, 1993; Van Gog & Paas, 2008). The classroom experiment replicated this worked example efficiency and demonstrated that this was the case for novices, but not for learners with relatively more prior knowledge. Hence, these experiments were the first to show that the worked example effect also applies to novices' training of CT-skills (e.g., Paas & Van Gog, 2006; Renkl, 2014; Van Gog et al., 2019). The observation from the second (i.e., classroom) experiment also supports findings regarding the expertise reversal effect (e.g., Kalyuga et al., 2003, 2012), which shows that while instructional strategies that assist learners in developing cognitive schemata are effective for low-knowledge learners, they are often not effective for higher-knowledge learners.

The classroom study reported in **Chapter 4** investigated whether comparing correct and erroneous examples (i.e., contrasting examples) would enhance unbiased reasoning more than studying correct examples only, studying erroneous examples only, and solving practice problems. Students were provided with the CT-instructions and practice on domain-relevant problems, under one of the four conditions. Results revealed that students' learning outcomes again improved from pretest to posttest. Moreover, their performance improved even further after a three-week and nine-month delay, although the latter finding could also be attributed to the further instructions that were given in courses in-between the three-week and nine-month follow up. Unexpectedly, however, results did not reveal any differences among conditions on either learning outcomes or transfer performance and, thus, differ from findings of previous studies (e.g., Durkin & Rittle-johnson, 2012; Kawasaki, 2010; Loibl & Leuders, 2018, 2019; Siegler, 2002). Moreover, it is surprising that this study did not reveal a beneficial effect of studying correct examples as opposed to practicing with problems (i.e., worked example effect), which is contrary to the finding in Chapter 3 and findings of previous studies on many other tasks (e.g., Renkl, 2014; Van Gog et al., 2019).

In **Chapter 5**, a classroom study was described that empirically investigated whether repeated retrieval practice over time (i.e., working on practice tasks in sessions that were weeks apart), would be beneficial for learning to reason in an unbiased manner and whether it can additionally facilitate transfer. Students were instructed on CT and avoiding belief-bias in syllogistic reasoning and practiced with syllogisms on domain-relevant problems. After each practice-task, they received correct-answer feedback and were given a worked example. Depending on assigned condition, they did not engage in extra practice, practiced a second time (week later), or practiced a second (week later) *and* third time (two weeks after second time). Consistent with previous repeated retrieval findings (e.g., Butler, 2010; McDaniel et al., 2012, 2013; Roediger & Butler, 2011), results revealed that average performance scores *during* practice sessions increased with more repetitions. However, repeated retrieval practice did not have a significant effect, compared to practicing just once, on learning outcomes on the final test, as judged by total scores (MC-answers plus justification). Exploring performance on MC-answers only revealed pretest to posttest learning gains, suggesting that students did benefit from instruction/practice but may have been unable to justify their answers. Effects on transfer could not be tested due to a floor effect. It seems possible that the feedback after each practice task eliminated the effects of repeated retrieval, in line with findings from recent research (Kliegl et al., 2019; Pastötter & Bäuml, 2016; Storm et al., 2014), since students spent more time on worked-example feedback after incorrect than correct retrievals.

The study in **Chapter 6** focused exclusively on identifying whether unsuccessful transfer of CT-skills would be due to a failure to recognize that acquired knowledge is relevant in a new context, to recall that knowledge, or to apply that knowledge to the new context (i.e., the three-step model of transfer; Barnett & Ceci, 2012). In two experiments (classroom and laboratory), students received explicit instructions on CT and avoiding belief-bias in syllogistic reasoning and practiced with syllogisms on domain-relevant problems. This time, students' performance was measured on syllogisms with different story contexts (to assess learning), syllogisms in a different format (to assess near transfer), and novel tasks that shared similar features with syllogisms (to assess far transfer) both on a pretest and immediate posttest. On the posttest transfer items, students received no support, received hints that the information provided in the learning phase is relevant for these items (recognition support), received hints that the information provided in the learning phase is relevant and were prompted to recall the acquired knowledge (free recall), or received hints that the information provided in the learning phase is relevant and receiving a reminder of the paper-based overview of that information that they received (recall support). The effects of support for different steps in the process were compared to infer where difficulties arise for learners (cf. Butler et al, 2013, 2017). Additionally, it was explored (within the free recall condition) whether students' ability to recall the acquired knowledge was related to their posttest performance on near and far transfer items. Over the two experiments, learning and near transfer outcomes improved after instruction/practice (i.e., from pretest to posttest). Results even showed some increase on far transfer items, but the far transfer scores were overall rather low, so there was still a lot of room for improvement. Interestingly, students did not benefit from recognition and recall support while solving transfer tasks (i.e., there were no significant differences among conditions). This finding suggests that students were able to recognize that the acquired knowledge was relevant to the new task and to recall that knowledge, but had difficulties in applying the relevant knowledge to the new tasks. However, findings from the free recall condition do not fully support the idea that it is only an application/mapping problem. Most students did not retrieve all relevant information from memory, and exploratory analyses pointed to moderate-to-large positive correlations between students' retrieved knowledge and their performance on near and far transfer items. This may suggest that suboptimal recall is at least partially responsible for unsuccessful transfer as well. Descriptive statistics support this idea: students who received recall support had higher (though not significantly higher) scores than the other conditions on far transfer items at posttest in the laboratory study and on near transfer items at posttest in the classroom study.

Discussion of main findings

Together, the studies in this dissertation seem to corroborate findings of previous studies on teaching CT in general (Abrami et al., 2014, 2018) and unbiased reasoning in particular (Heijltjes et al., 2014a, 2014b, 2015) that providing students with explicit CT-instruction and the opportunity to practice with domain-relevant problems improves learning outcomes. Although we did not include a no-instruction/practice control condition, students did show pretest to posttest performance gains on practiced/instructed items, and their performance remained stable or improved even further after a delay of (several) weeks (Chapters 2 to 4). Regarding the effect of instruction/practice on transfer, the study in Chapter 6 showed a noticeable progress on near transfer from pretest to posttest, that is, after instructions and practice activities[12]. However, there were no or very limited indications of progress on far transfer (Chapters 2 and 6, respectively)[13], which is in line with findings of previous studies that examined effects on *far* transfer (Heijltjes et al., 2014a, 2014b, 2015). Taken together, this research extends prior research on teaching for transfer of CT-skills and confirms that transfer between closely related situations occurred more often than transfer between situations that had less in common (Barnett & Ceci, 2002; Bray, 1928; Dinsmore et al., 2014).

Remarkably, the generative processing strategies did not work as expected: Chapters 2 to 5 found no indications that these strategies – be it self-explaining during practice, interleaved practice, comparing correct and erroneous examples, or repeated retrieval practice – further improved learning or transfer of CT-skills. It has been well established that encouraging generative processing fosters knowledge acquisition and transfer of various cognitive skills (e.g., Fiorella & Mayer, 2016; Wittrock, 2010). As such, it is somewhat surprising that generative processing strategies did not seem beneficial for fostering CT-skills.

There are several possible explanations for this absence of differential effects of generative processing strategies on learning and transfer. The possible strategy-specific explanations and preconditions have been addressed in the respective chapters, so I will not repeat them here, but instead, I will focus on the overarching issues. First, it seems possible that the CT-instructions, which included worked examples, already had a substantial effect on learning unbiased reasoning, making it difficult to find differential effects of different types of practice activities. Most studies on the effects of generative processing strategies with other types of cognitive tasks use pure practice conditions or

[12] Near transfer items were only included in the tests of the study presented in Chapter 6.

[13] The studies presented in Chapters 3, 4, and 5 did not include transfer items in the pretest and were, therefore, not able to detect transfer gains.

give minimal instructions prior to practice (e.g., Fiorella & Mayer, 2016). Thus, the effects are usually not investigated in a context in which elaborate processing of instructions precedes practice, as in the studies in this dissertation.

Second, the absence of differential effects of generative processing on learning may be related to the affective and attitudinal dimension of CT. Being able to think critically relies on the extent to which one possesses the requisite skills and is able to use these skills, but also on whether one is inclined to use these skills (i.e., thinking dispositions; Perkins et al., 1983). It is possible, for instance, that generative processing would only benefit students who score high on thinking dispositions (such as need for cognition, Cacioppo & Petty, 1982, or actively open-minded thinking, Stanovich & West, 2007). A possible interaction between generative processing strategies and thinking dispositions could not be investigated in this dissertation, however, because thinking dispositions were not assessed.

Third, in some studies, the classroom setting might explain why there were no differential effects of generative processing. In Chapter 2, I already pointed to the possibility that because the study was conducted in an existing CT-course (as in all classroom studies part of this dissertation), students' willingness to invest effort in their performance may have been higher than generally in psychological laboratory studies. The learning materials from the study were also relevant for the course/exam and their performance actually mattered (intrinsically or extrinsically) to them. Not so much on the posttest of this study, which did not have consequences for their exam grade, but on such tasks in general. As such, students in the control condition may have engaged in generative processing themselves, for instance by covertly trying to come up with explanations for the questions. It is therefore possible that effects of generative processing strategies such as self-explaining found in the psychological laboratory – where students participate to earn required research credits and the learning materials are not part of their study program – might not readily transfer to field experiments conducted in real classrooms. This could be a possible explanation for the lack of effects of contrasting examples (Chapter 4) as well, in which the control conditions may have tried to compare the given correct (or erroneous) examples with internally represented erroneous (or correct) solutions. I will discuss recommendations for future research based on this assumption later in this chapter. It should be noted though, that the above argument probably cannot fully explain the absence of differential effects of interleaved practice (Chapter 3) and repeated retrieval practice (Chapter 5), where motivational aspects are less crucial. To illustrate, in Chapter 3, students in the control condition practiced in a blocked schedule and could not easily engage in interleaved practice themselves. Moreover, Chapter 3 included both a classroom *and* laboratory study and consistently demonstrated a lack of differential effects and, therefore, the classroom setting argument cannot fully explain the absence of differential effects of this generative processing strategy.

A possible reason for the lack of transfer to novel problem types in general, might be related to the duration or extensiveness of the practice activities. Even though substantial evidence is provided that students learned to solve abstract heuristics-and-biases tasks (Chapters 2 to 6) and tasks closely related to those instructed (Chapter 6), their subject-matter knowledge may have been insufficient for solving more complex or novel CT-tasks. That might explain the considerably low levels of performance on far transfer items in all chapters. As such, it can be argued that establishing transfer to novel problem types needs longer or more extensive practice. Additionally, Chapter 6 implies that instructional interventions aimed at far transfer of CT-skills should focus on recall of the acquired knowledge and application of that knowledge onto novel tasks, since students seem to have most difficulty with these steps in the transfer process (for the three-step model of transfer, see Barnett & Ceci, 2012). These explanations are not mutually exclusive and should be investigated further in future research. I will elaborate on this when giving suggestions for future work. Nonetheless, the series of studies presented in this dissertation do show – contrary to the assertion made by Halpern and Butler (2019) that teaching CT-skills explicitly with multiple examples from different contexts will facilitate transfer to novel contexts – that establishing transfer of CT-skills to novel *problem types* is no easy feat, at least with regard to skills required for unbiased reasoning.

Taken together, providing students with explicit CT-instruction and opportunities to practice with domain-relevant problems is beneficial for learning unbiased reasoning, but what kind of practice activity does not seem to matter. The latter finding may be explained by the magnitude of the effect of the CT-instruction itself, the nature of the practice tasks (i.e., heuristics-and-biases tasks), and/or the setting of the experiments. Furthermore, findings suggest that these instructions and practice opportunities may also enhance near transfer, but are not sufficient to establish further transfer. As such, it can be suggested that bringing about far transfer needs longer or more extensive practice, in which obstacles such as suboptimal recall and application should be countered.

Methodological issues

Several methodological issues need to be discussed. Again, I will focus on the overarching issues as study specific issues have been addressed in each chapter. First, the measures in Chapters 2 to 4 showed low levels of reliability. Reliability issues are quite common in research using tests consisting of heuristics-and-biases tasks (Aczel et al., 2015b; Bruine de Bruin, 2007; Janssen et al., 2019a; West et al., 2008) and multiple studies revealed concerns with the reliability of widely used standardized CT tests, particularly with regard to subscales (Bernard et al., 2008; Bondy et al., 2001; Janssen et al., 2020; Ku, 2009; Liu et al., 2014; Leppa, 1997; Loo & Thorpe, 1999; Rear, 2019). Low levels of reliability decrease statistical power and, thereby, reduce the chance of detecting true effects (e.g., Cleary et al., 1970; Rogers & Hopkins, 1988). Furthermore, given that the point estimates of the crucial interaction effects appeared to be very small, these may have been difficult to detect.

In this dissertation, the low levels of reliability can probably be explained in terms of multidimensionality of the tests encompassing several heuristics-and-biases tasks, a factor often ignored in current research. That is, when tests represent multiple constructs that do not correlate with each other. As alluded to earlier, performance on such tasks depends not only on the extent to which that task elicits a bias (resulting from heuristic reasoning), but also on the extent to which one possesses the requisite mindware. Thus, systematic variance in performance on such tasks can either be explained by a person's use of heuristics or his/her available mindware. If it differs per item to what extent a correct answer depends on these two aspects, there may not be a common factor explaining all interrelationships between the measured items. In that case, the theoretical assumption of unidimensionality is violated.

In the research presented in this dissertation, the general reliability issue may have increased even more since multiple task types were included in the CT-skills tests, requiring different types of mindware (e.g., rules of logic or probability). Hence, I have attempted to increase reliability of the measures in Chapters 5 and 6, by constructing tests with multiple items of one task category to narrow down the tests into single measurable constructs and, thereby, to decrease measurement error (LeBel & Paunonen, 2011). Indeed, these compositions led to quite reliable measures. However, even though biased reasoning is a very important aspect of CT, it is already a rather restricted operationalization and this focus on one task category narrowed it even further. To achieve further progress in research on instructional methods for teaching CT, more knowledge on the construct validity of CT in general and unbiased reasoning is needed, and reliable (aspect-specific) tests of CT should be developed. That seems challenging, however, especially given that for practical use in educational contexts, tests cannot be overly long.

Along with the issues raised, it should be considered to what extent the tests in the research reported in this dissertation and previous research by others, accurately assessed CT as it is practised in the real world, that is, outside education. I would argue that the current findings do provide valuable insights into how people reason, given that heuristics-and-biases tasks represent how people judge under uncertainty and in various contexts; heuristics and biases appear in newspapers, books, courses, and applications of many kinds. Especially since in this dissertation – contrary to standardized CT-tests and most research on heuristics-and-biases tasks – CT was assessed at the level of individual study domains (i.e., content of the tasks was adapted to specific study domains) and could, therefore, be evaluated within authentic contexts. To illustrate, in Chapter 6, students' ability to evaluate the logical validity of arguments in a written news item or article on a topic that they might encounter in their working life, was assessed. Hence, performance on these tasks could presumably predict everyday reasoning, as has already been assumed in various studies (see for example, Gilovich et al., 2002).

A strength of the research presented in this dissertation, is that it follows the standards of an open research culture by using open practices. Open practices are designed to make scientific processes and results more transparent and accessible to others than the researchers involved (e.g., Nosek et al., 2015). It includes making complete research materials, designs, and data freely available to anyone, which makes it easier to replicate and evaluate scientific findings (for instance because both null results and statistically significant results are accessible). Although transparency and openness are readily recognized as disciplinary norms and values, scientific practice often fails to adhere these valued features (Ioannidis et al., 2014; John et al., 2012; Open Science Collaboration, 2012). Practicing open science has been central to the research reported in this dissertation. The study presented in Chapter 2 has already been published open access[14] and, for all studies, important aspects of the research design and data analyses are publicly available on the online repository 'Open Science Framework' (OSF).

Furthermore, this dissertation is strengthened by the fact that some of the studies have been preregistered on the OSF repository (Chapters 5 and 6), with specific details such as the hypotheses, planned analyses, and rules for data exclusions recorded prior to the data-analyses. The practice of pre-registration was introduced in response to some serious issues in academic publishing. These included, for instance, the use of 'questionable research practices' by individual researchers, such as manipulating statistics to obtain significant effects (*p*-hacking) and hypothesizing after the results are known (HARKing; John et al., 2012; Simmons et al., 2011). Both open practices and pre-registrations help to more accurately assess the evidence base for phenomena and are, therefore, imperative to increase confidence in scientific findings.

[14] In time, the studies presented in the other chapters will be publicly available as well, through publications in open access journals or preprints on the OSF-repository.

Implications for practice and future directions

Educational practice and future research could benefit from the findings presented in this dissertation both from a theoretical and practical point of view. The findings clearly indicate that providing students with explicit CT-instruction and the opportunity to practice with domain-relevant problems has the potential to improve learning. It is important to emphasize, however, that there is no one-size-fits-all recommendation in terms of best practice activity. These acquired insights advocate for CT integration in higher education curricula and explicit CT objectives at course level (i.e., CT as important learning outcome), for instance through explicit CT-courses. Acquisition of requisite mindware was particularly central to this dissertation, but perhaps instructional designs should pay more attention to changing students' thinking dispositions. To illustrate, a student who masters CT-skills but is unwilling to put in the mental effort to use these skills on complex or novel CT-tasks, will be no better off than a student without these CT-skills. Investigating the exact role of students' thinking dispositions in fostering unbiased reasoning and developing an approach aimed at improving *both* aspects will be a fruitful area for further work. Enhancing thinking dispositions may require building a certain culture of thinking in the classroom, in which students are exposed to models of thinking of fellow students, supported in cultural interaction, and provided with direct instructions on thinking dispositions (cf. enculturation model; Tishman et al., 1993). Also, future research could explore whether changes as complex as these may be realized through personalized approaches, such as personalized feedback (Marsh & Eliseev, 2019). It is important, then, to ensure that students believe in and process that feedback (Rich et al., 2017), which, however, is often left to the discretion of students.

To specifically address development of deep learning of CT-skills, instructional design studies as in this dissertation could be preceded by research that identifies the exact factors that help or hinder learning of that explicit CT-skill. Chapter 6 is a useful example of how to design such a study. This chapter provided initial insights into the obstacles that prevent successful transfer of overturning belief-biased responses when evaluating the logical validity of arguments; students seem to have most difficulty with recall of the acquired knowledge and application of that knowledge onto novel tasks. Future studies should therefore focus on the recall and application/mapping steps in the transfer process. However, it could not be determined from this study *why* students have difficulties with these steps in the transfer process, which should be addressed in future investigations. Furthermore, the question of how to facilitate transfer of CT-skills remains of interest. Assuming that unsuccessful transfer of CT-skills can be attributed to recall and application/mapping problems, the challenge for researchers and educational practitioners (e.g., consultants, teachers) in the CT-domain is to develop instructional designs that focus on these steps in the transfer process. A possible direction could be to provide exemplars of knowledge application while gradually remove scaffolding (cf. four-component instructional design model; Van Merriënboer et al., 1992) or while fading from concrete-to-abstract situations (i.e., concreteness fading; McNeil & Fyfe, 2012).

More broadly, a key challenge of classroom studies is to prevent noisy or incomplete data produced by these realistic settings (e.g., Hulleman & Cordray, 2009), which would make it more difficult to detect any (small) effect. Issues as these can be (at least partially) addressed by using large sample sizes and collecting multiple data points per participant. Moreover, to increase the impact, transfer, and translation of education research into improved practice, it seems promising to additionally conduct instructional design research within even more realistic settings than in this dissertation (e.g., through education design research; McKenney & Reeves, 2018). Education design research blends empirical investigation with systematic development and implementation of solutions, such as improved instructional designs, for educational problems. To establish transfer of CT-skills, a longer, but carefully structured, intervention based on principles derived from prior fundamental research may be needed. A comprehensive CT-course (that fosters the cultivation of both CT-skills and thinking dispositions) can possibly meet these needs, which takes long-term studies in realistic settings to test its effectiveness.

All of this assumes, of course, that those who teach CT are equipped with the knowledge and skills needed to effectively teach unbiased reasoning (e.g., Elen et al., 2019; Klassen & Tze, 2014). The challenge is for educators to know what is needed, and when. Furthermore, for educators to teach CT, they need to consider teaching CT as relevant (e.g., Eccles & Wigfield, 2002; Elen et al., 2009) and should have confidence in their ability to teach CT (Janssen et al., 2019b). To achieve this ambitious goal, we can facilitate educators by including (teaching/explaining) CT in professional development programs (e.g., Janssen et al., 2019a) and sharing CT resources that they could use in their own courses (e.g., Dutch online platform Kritisch Leren Denken: https://kritischdenkenhbo.nl/).

Conclusion

This dissertation sheds light on fostering higher education students' learning and transfer of CT-skills, focusing specifically on avoiding bias in reasoning. The evidence presented highlights the importance of explicit CT-instruction and practice opportunities for learning of these skills. It also demonstrated that generative processing is not a panacea for all kinds of learning tasks: it does not seem to improve learning and transfer of CT-skills required for unbiased reasoning. This dissertation again underlines the great difficulty encountered when seeking to enhance CT-skills in such a way that these would also transfer across tasks/domains. All things considered, to help students become good critical thinkers in the sense they can apply the acquired skills to a variety of tasks and contexts, it seems valuable to develop longer CT interventions or comprehensive CT courses. To conclude, I believe further progress in this area will come from instruction designs that are grounded in solid laboratory and classroom studies.

References

Abel, M., & Roediger, H. L. (2017). Comparing the testing effect under blocked and mixed practice: The mnemonic benefits of retrieval practice are not affected by practice format. *Memory & Cognition, 45*, 81–92. https://doi.org/10.3758/s13421-016-0641-8

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*, 1102–1134. https://doi.org/10.3102/0034654308326084

Abrami, P. C., Bernard, R. M., Borokhovski, E., Waddington, D. I., Wade, C. A., & Persson, T. (2014). Strategies for teaching students to think critically: A Meta-Analysis. *Review of Educational Research, 85, 275–314*. https://doi.org/10.3102%2F0034654314551063

Aczel, B., Bago, B., Szollosi, A., Foldes, A., & Lukacs, B. (2015a). Is it time for studying real-life debiasing? Evaluation of the effectiveness of an analogical intervention technique. *Frontiers in psychology, 6*, 1120. https://doi.org/10.3389/fpsyg.2015.01120

Aczel, B., Bago, B., Szollosi, A., Foldes, A., & Lukacs, B. (2015b). Measuring individual differences in decision biases: Methodological considerations. *Frontiers in Psychology, 6*, 1770. https://doi.org/10.3389/fpsyg.2015.01770

Adams, D. M., McLaren, B. M., Durkin, K., Mayer, R. E., Rittle-Johnson, B., Isotani, S., & Van Velsen, M. (2014). Using erroneous examples to improve mathematics learning with a web-based tutoring system. *Computers in Human Behavior, 36*, 401–411. https://doi.org/10.1016/j.chb.2014.03.053

Ajayi, T., & Okudo, J. (2016). Cardiac arrest and gastrointestinal bleeding: A case of medical heuristics. *Case Reports in Medicine, 2016*. https://doi.org/10.1155/2016/9621390

Albaret, J. M., & Thon, B. (1998). Differential effects of task complexity on contextual interference in a drawing task. *Acta Psychologica, 100*, 9–24. https://doi.org/10.1016/S0001-6918(98)00022-5

Ananiadou, K., & Claro, M. (2009). *21st Century skills and competences for new millennium learners in OECD countries*. OECD Publishing. https://dx.doi.org/10.1787/218525261154

Angeli, C., & Valanides, N. (2009). Instructional effects on critical thinking: Performance on ill-defined issues. *Learning and Instruction, 19*, 322–334. http://doi.org/10.1016/j.learninstruc.2008.06.010

Arghami, N. R., & Billard, L. (1982). A modification of a truncated partial sequential procedure. *Biometrika, 69*, 613–618. https://doi.org/10.1093/biomet/69.3.613

Argharni, N. R., & Billard, L. (1991). A partial sequential t-test. *Sequential Analysis, 10*, 181–197.

Aron, A. R. (2008). Progress in executive-function research: From tasks to functions to regions to networks. *Current Directions in Psychological Science, 17*, 124–129. https://doi.org/10.1111%2Fj.1467-8721.2008.00561.x

Arum, R., & Roksa, J. (2011). Limited learning on college campuses. *Society, 48*, 203–207. https://doi.org/10.1007/s12115-011-9417-8

Atkinson, R. K., Renkl, A., & Merrill, M. M. (2003). Transitioning from studying examples to solving problems: Effects of self-explanation prompts and fading worked-out steps. *Journal of Educational Psychology, 95*, 774–783. https://doi.org/10.1037/0022-0663.95.4.774

Avans Hogeschool (2014a). *Ambitie 2020: Het verschil maken.*

Avans Hogeschool (2014b). *Onderwijsvisie: Samen het maximale uit jezelf halen.*

Avans Hogeschool (2019). *Routekaart Ambitie 2025.*

Avans Hogeschool (n.d.). *Platform Kritisch leren denken.* Retrieved January 31, 2020, from www.kritischdenkenhbo.nl

Bahrick, H. P. (1979). Maintenance of knowledge: Questions about memory we forgot to ask. *Journal of Experimental Psychology: General, 108*, 296–308. https://psycnet.apa.org/doi/10.1037/0096-3445.108.3.296

Barbieri, C., & Booth, J. L. (2016). Support for struggling students in algebra: Contributions of incorrect worked examples. *Learning and Individual Differences, 48*, 36–44. https://doi.org/10.1016/j.lindif.2016.04.001

Barnett, S. M., & Ceci, S. J. (2002). When and where do we apply what we learn?: A taxonomy for far transfer. *Psychological Bulletin, 128*, 612–636. http://dx.doi.org/10.1037/0033-2909.128.4.612

Barreiros, J., Figueiredo, T., & Godinho, M. (2007). The contextual interference effect in applied settings. *European Physical Education Review, 13*, 195–208.

Baron, J. (2008). *Thinking and deciding* (4th ed). Cambridge University Press.

Barreiros, J., Figueiredo, T., & Godinho, M. (2007). The contextual interference effect in applied settings. *European Physical Education Review, 13*, 195–208. https://doi.org/10.1177/1356336X07076876

Bassok, M., & Holyoak, K. J. (1989). Interdomain transfer between isomorphic topics in algebra and physics. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15*, 153–166. http://dx.doi.org/10.1037/0278-7393.15.1.153

Battig, W. F. (1978). The flexibility of human memory. In L. S. Cermak & F. I. M. Craik (Eds.), *Levels of processing and human memory*. Erlbaum.

Benjamin, A. S., & Tullis, J. (2010). What makes distributed practice effective?. *Cognitive Psychology, 61*, 228–247. https://doi.org/10.1016/j.cogpsych.2010.05.004

Bernard, R. M., Zhang, D., Abrami, P. C., Sicoly, F., Borokhovski, E., & Surkes, M. A. (2008). Exploring the structure of the Watson–Glaser Critical Thinking Appraisal: One scale or many subscales? *Thinking Skills and Creativity, 3*, 15–22. https://doi.org/10.1016/j.tsc.2007.11.001

Berry, D. C. (1983). Metacognitive experience and transfer of logical reasoning. *The Quarterly Journal of Experimental Psychology, 35*, 39–49. https://doi.org/10.1080/14640748308402115

Berthold, K., Eysink, T. H. S., & Renkl, A. (2009). Assisting self-explanation prompts are more effective than open prompts when learning with multiple representations. *Instructional Science, 37*, 345–363. https://doi.org/10.1007/s11251-008-9051-z

Bertsch, S., Pesta, B. J., Wiscott, R., & McDaniel, M. A. (2007). The generation effect: A meta-analytic review. *Memory & cognition, 35*, 201–210. https://doi.org/10.3758/BF03193441

Best, J. R., Miller, P. H., & Jones, L. L. (2009). Executive functions after age 5: Changes and correlates. *Developmental Review, 29*, 180–200. https://doi.org/10.1016/j.dr.2009.05.002

Beyer, B. (2008). How to teach thinking skills in social studies and history. *The Social Studies, 99*, 196-201. http://dx.doi.org/10.3200/TSSS.99.5.196-201

Billings, L., & Roberts, T. (2014). *Teaching critical thinking: Using seminars for 21st century literacy*. Routledge.

Bisra, K., Liu, Q., Nesbit, J. C., Salimi, F., & Winne, P. H. (2018). Inducing self-explanation: A meta-analysis. *Educational Psychology Review, 30*, 703–725. https://doi.org/s10648-018-9434-x

Bjork, R.A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe and A. Shimamura (Eds.), *Metacognition: Knowing about knowing*. (pp. 185–205). MIT Press.

Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In M. A. Gernsbacher, R. W. Pew, & J. R. Pomerantz (Eds.), *Psychology and the real world: Essays illustrating fundamental contributions to society* (pp. 59–68). Worth Publishers.

Bjork, E., Soderstrom, N., & Little, J. (2015). Can multiple-choice testing induce desirable difficulties? Evidence from the laboratory and the classroom. *The American Journal of Psychology, 128*, 229–239. https://doi.org/10.5406/amerjpsyc.128.2.0229

Bondy, K. N., Koenigseder, L. A., Ishee, J. H., & Williams, B. G. (2001). Psychometric properties of the California Critical Thinking Tests. *Journal of Nursing Measurement, 9*, 309–328. https://doi.org/10.1891/1061-3749.9.3.309

Booth, J. L., Lange, K. E., Koedinger, K. R., & Newton, K. J. (2013). Using example problems to improve student learning in algebra: Differentiating between correct and incorrect examples. *Learning and Instruction, 25*, 24–34. https://doi.org/10.1016/j.learninstruc.2009.10.001

Booth, J. L., Oyer, M. H., Paré-Blagoev, E. J., Elliot, A. J., Barbieri, C., Augustine, A., & Koedinger, K. R. (2015). Learning algebra by example in real-world classrooms. *Journal of Research on Educational Effectiveness, 8*, 530–551. https://doi.org10.1080/19345747.2015.1055636

Botella, J., Ximénez, C., Revuelta, J., & Suero, M. (2006). Optimization of sample size in controlled experiments: The CLAST rule. *Behavior Research Methods, Instruments, & Computers, 38*, 65–76. https://doi.org/10.3758/BF03192751

Bray, C. W. (1928). Transfer of learning. *Journal of Experimental Psychology, 11*, 443–467. http://dx.doi.org/10.1037/h0071273

Bruine de Bruin, W., Parker, A. M., & Fischhoff, B. (2007). Individual differences in adult decision-making competence. *Journal of Personality and Social Psychology, 92*, 938–956. https://doi.org/10.1037/0022-3514.92.5.938

Brunmair, M., & Richter, T. (2019). Similarity matters: A meta-analysis of interleaved learning and its moderators. *Psychological Bulletin, 145*, 1029. https://psycnet.apa.org/doi/10.1037/bul0000209

Butler, A. C. (2010). Repeated testing produces superior transfer of learning relative to repeated studying. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*, 1118–1133. http://dx.doi.org/10.1037/a0019902

Butler, A. C., Black-Maier, A. C., Raley, N. D., & Marsh, E. J. (2017). Retrieving and applying knowledge to different examples promotes transfer of learning. *Journal of Experimental Psychology: Applied, 23*, 433–446. https://doi.org/10.1037/xap0000142

Butler, A. C., Godbole, N., & Marsh, E. J. (2013). Explanation feedback is better than correct answer feedback for promoting transfer of learning. *Journal of Educational Psychology, 105*, 290–298. https://psycnet.apa.org/doi/10.1037/a0031026

Butler, H. A., & Halpern, D. F. (2020). Critical thinking impacts our everyday lives. In Sternberg R. J., & Halpern, D. F. (Eds.), *Critical Thinking in Psychology* (pp. 152–162). Cambridge University Press.

Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology, 42*, 116–131. https://doi.org/10.1037/0022-3514.42.1.116

Carpenter, S. K. (2012). Testing enhances the transfer of learning. *Current Directions in Psychological Science, 21*, 279–283. https://doi.org/10.1177/0963721412452728

Carpenter, S. K., & Kelly, J. W. (2012). Tests enhance retention and transfer of spatial learning. *Psychonomic Bulletin & Review, 19*, 443–448. https://doi.org/10.3758/s13423-012-0221-2

Carpenter, S. K., & Mueller, F. E. (2013). The effects of interleaving versus blocking on foreign language pronunciation learning. *Memory & Cognition, 41*, 671–682. https://doi.org/10.3758/s13421-012-0291-4

Catapano, R., Tormala, Z. L., & Rucker, D. D. (2019). Perspective taking and self-persuasion: Why "putting yourself in their shoes" reduces openness to attitude change. *Psychological Science, 30*, 1–12. Htttps://doi.org/10.1177/0956797618822697

Cepeda, N. J., Vul, E., Rohrer, D., Wixted, J. T., & Pashler, H. (2008). Spacing effects in learning: A temporal ridgeline of optimal retention. *Psychological Science, 19*, 1095–1102. https://doi.org/10.1111/j.1467-9280.2008.02209.x

Chapman, L. J., & Chapman, J. P. (1967). Genesis of popular but erroneous psychodiagnostic observations. *Journal of Abnormal Psychology, 72*, 193–204. https://psycnet.apa.org/doi/10.1037/h0024670

Charter, R. A. (2003). Study samples are too small to produce sufficiently precise reliability coefficients. *The Journal of General Psychology, 130*, 117–129. https://doi.org/10.1080/00221300309601280

Chi, M. (2000). Self-explaining expository texts: The dual processes of generating inferences and repairing mental models. *Advances in Instructional Psychology, 5*, 161–238.

Chi, M. T. H., de Leeuw, N., Chiu, M., & LaVancher, C. (1994). Eliciting self-explanation improves understanding. *Cognitive Science, 18*, 439–477. https://doi.org/10.1207/s15516709cog1803_3

Cleary, T. A., Linn, R. L., & Walster, G. W. (1970). Effect of reliability and validity on power of statistical tests. *Sociological Methodology, 2*, 130–138. https://doi.org/10.1037/a0031026

Cohen, J (1988). *Statistical power analysis for the behavioral sciences* (2nd. ed., reprint). Psychology Press.

Cormier, S. M., & Hagman, J. D. (Eds.). (2014). *Transfer of learning: Contemporary research and applications*. Academic Press.

Cull, W.L. (2000). Untangling the benefits of multiple study opportunities and repeated testing for cued recall. *Applied Cognitive Psychology, 14*, 215–235. https://doi.org/10.1002/(SICI)1099-0720(200005/06)14:3<215::AID-ACP640>3.0.CO;2-1

Darling-Hammond, L. (2010). Teacher education and the American future. *Journal of Teacher Education, 61*, 35–47. https://doi.org/10.1177/0022487109348024

Davies, M. (2013). Critical thinking and the disciplines reconsidered. *Higher Education Research & Development, 32*, 529–544. https://doi.org/10.1080/07294360.2012.697878

Dawson, E., Gilovich, T., & Regan, D. T. (2002). Motivated reasoning and performance on the Wason selection task. *Personality and Social Psychology Bulletin, 28*, 1379–1387. https://doi.org/10.1177/014616702236869

DeAngelo, L., Hurtado, S., Pryor, J. H., Kelly, K. R., Santons, J. L., Korn, W. S. (2009). *The American college teacher: National norms for the 2007-2008 HERI faculty survey*. Higher Education Research Institute.

De Bruin, A. B., Rikers, R. M., & Schmidt, H. G. (2007). The effect of self-explanation and prediction on the development of principled understanding of chess in novices. *Contemporary Educational Psychology, 32*, 188–205. https://doi.org/10.1016/j.cedpsych.2006.01.001

De Chantal, P. L., Newman, I. R., Thompson, V., & Markovits, H. (2019). Who resists belief-biased inferences? The role of individual differences in reasoning strategies, working memory, and attentional focus. *Memory & Cognition, 48*, 655–671. https://doi.org/10.3758/s13421-019-00998-2

De Croock, M. B., & van Merriënboer, J. J. (2007). Paradoxical effects of information presentation formats and contextual interference on transfer of a complex cognitive skill. *Computers in Human Behavior, 23*, 1740–1761. https://doi.org/10.1016/j.chb.2005.10.003

De Croock, M. B., van Merriënboer, J. J., & Paas, F. G. (1998). High versus low contextual interference in simulation-based training of troubleshooting skills: Effects on transfer performance and invested mental effort. *Computers in Human Behavior, 14*, 249–267. https://doi.org/10.1016/S0747-5632(98)00005-3

De Koning, B. B., Tabbers, H. K., Rikers, R. M. J. P., & Paas, F. (2010). Learning by generating vs. receiving instructional explanations: Two approaches to enhance attention cueing in animations. *Computers & Education, 55*, 681–691. https://doi.org/10.1016/j.compedu.2010.02.027

Delaney, P. F., Verkoeijen, P. P. J. L., & Spirgel, A. (2010). Spacing and testing effects: A deeply critical, lengthy, and at times discursive review of the literature. In B. H. Ross (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 53, pp. 63–147). Academic Press. https://doi.org/10.1016/S0079-7421(10)53003-2

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

DeWinstanley, P. A., & Bjork, E. L. (2004). Processing strategies and the generation effect: Implications for making a better reader. *Memory & Cognition, 32*, 945–955. https://doi.org/10.3758/BF03196872

Dinsmore, D. L., Baggetta, P., Doyle, S., & Loughlin, S. M. (2014). The role of initial learning, problem features, prior knowledge, and pattern recognition on transfer success. *The Journal of Experimental Education, 82*, 121–141. https://doi.org/10.1080/00220973.2013.835299

Doll, R. (1982). Clinical trials: Retrospect and prospect. *Statistics in Medicine, 1,* 337–344. https://doi.org/10.1002/sim.4780010411

Druckman, D., & Bjork, R. A. (1994). *Learning, remembering, believing: Enhancing human performance.* National Academy Press.

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest, 14*, 4–58. https://doi.org/10.1177/1529100612453266

Durkin, K., & Rittle-Johnson, B. (2012). The effectiveness of using incorrect examples to support learning about decimal magnitude. *Learning and Instruction, 22*, 206–214. https://doi.org/10.1016/j.learninstruc.2011.11.001

Duron, R., Limbach, B., & Waugh, W. (2006). Critical thinking framework for any discipline. *International Journal of Teaching and Learning in Higher Education, 17*, 160–166. http://doi.org/10.1016/j.nepr.2006.09.004

Ebbinghaus, H. (1964). *Memory: A contribution to experimental psychology* (H. A. Ruger & C. E. Bussenius, Trans.). Dover (original work published 1885).

Eccles, J. S., & Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annual Review of Psychology, 53*, 109–132. https://doi.org/10.1146/annurev.psych.53.100901.135153.

Elen, J., Jiang, L., Huyghe, S., Evers, M., Verburgh, A., … (2019). *Promoting critical thinking in European higher education institutions: Towards an educational protocol*. C. Dominguez & R. Payan-Carreira (Eds.). Vila Real: UTAD.

Elia, F., Apra, F., Verhovez, A., & Crupi, V. (2016). "First, know thyself": Cognition and error in medicine. *Acta Diabetologica, 53*, 169–175. https://doi.org/10.1007/s00592-015-0762-8

Ennis, R. H. (1962). A concept of critical thinking. *Harvard Educational Review, 32*, 81–111.

Ennis, R. H. (1989). Critical thinking and subject specificity: Clarification and needed research. *Educational Researcher, 18*, 4–10. https://doi.org/10.3102/0013189X018003004

Ennis, R. (1992). The degree to which critical thinking is subject specific: clarification and needed research. In S. Norris (Ed.), *The generalizability of critical thinking: Multiple perspectives on an educational ideal* (pp. 21–37). Teachers College Press.

Evans, J. S. B. (1977). Toward a statistical theory of reasoning. *Quarterly Journal of Experimental Psychology, 29*, 621–635.

Evans, J. S. B. (2002). Logic and human reasoning: An assessment of the deduction paradigm. *Psychological Bulletin, 128*, 978–996. https://doi.org/10.1037/0033-2909.128.6.978

Evans, J. S. B. (2003). In two minds: Dual-process accounts of reasoning. *Trends in Cognitive Sciences, 7*, 454–459. https://doi.org/10.1016/j.tics.2003.08.012

Evans, J. S. B. T. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology, 59*, 255–278. https://doi.org/10.1146/annurev.psych.59.103006.093629

Evans, J. S. B., Barston, J. L., & Pollard, P. (1983). On the conflict between logic and belief in syllogistic reasoning. *Memory & Cognition, 11*, 295–306. https://doi.org/10.3758/BF03196976

Evans, J. S. B. (2011). Dual-process theories of reasoning: Contemporary issues and developmental applications. *Developmental Review, 31*, 86–102. https://doi.org/10.1016/j.dr.2011.07.007

Facione, P. A. (1990). *Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction*. The California Academic Press.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*, 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

Fiorella, L., & Mayer, R. E. (2015). *Learning as a generative activity*. Cambridge University Press.

Fiorella, L., & Mayer, R. E. (2016). Eight ways to promote generative learning. *Educational Psychology Review, 28*, 717–741. https://doi.org/10.1007/s10648-015-9348-0

Fischhoff, B. (1975). "Hindsight ≠ foresight: The effect of outcome knowledge on judgment under uncertainty". *Journal of Experimental Psychology: Human Perception and Performance*, 1, 288–299. https://doi.org/10.1037/0096-1523.1.3.288.

Fitts, D. A. (2010). Improved stopping rules for the design of efficient small-sample experiments in biomedical and biobehavioral research. *Behavior Research Methods, 42,* 3–22. https://doi.org/10.3758/BRM.42.1.3

Flores, K. L., Matkin, G. S., Burbach, M. E., Quinn, C. E., & Harding, H. (2012). Deficient critical thinking skills among college graduates: Implications for leadership. *Educational Philosophy and Theory, 44,* 212–230. https://doi.org/10.1111/j.1469-5812.2010.00672.x

Fong, G. T., Krantz, D. H., & Nisbett, R. E. (1986). The effects of statistical training on thinking about everyday problems. *Cognitive Psychology, 18*, 253–292. https://doi.org/10.1016/0010-0285(86)90001-0

Frick, R. W. (1998). A better stopping rule for conventional statistical tests. *Behavior Research Methods, Instruments, & Computers, 30*, 690–697. https://doi.org/10.3758/BF03209488

Fyfe, E. R., & Rittle-Johnson, B. (2017). Mathematics practice without feedback: A desirable difficulty in a classroom setting. *Instructional Science, 45*, 177–194. https://doi.org/10.1007/s11251-016-9401-1

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science, 7,* 155–170. https://doi.org/10.1016/S0364-0213(83)80009-3

Gentner, D. (1989). The mechanism of analogical learning. In Vosniadou, S., & Ortony, A. (Eds.), *Similarity and analogical reasoning* (pp. 119–241). Cambridge University Press.

Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology, 15,* 1–38. https://doi.org/10.1016/0010-0285(83)90002-6

Gigerenzer, G., & Hug, K. (1992). Domain-specific reasoning: Social contracts, cheating, and perspective change. *Cognition, 43*, 127–171. http://dx.doi.org/10.1016/0010-0277(92)90060-U

Gilovich, T., Griffin, D. & Kahneman D. (2002). *Heuristics and biases: The psychology of intuitive judgment*. Cambridge University Press.

Ginns, P. (2006). Integrating information: A meta-analysis of the spatial contiguity and temporal contiguity effects. *Learning and Instruction, 16*, 511–525. https://doi.org/10.1016/j.learninstruc.2006.10.001

Glaser, E.M. (1941). *An experiment in the development of critical thinking*. Teachers College. Columbia University.

Grabowski, B. (1996). Generative learning. Past, present, and future. In D. H. Jonassen (Ed.), *Handbook of research for educational communications and technology* (pp. 897–918). Macimillian Library Reference.

Greene, R. L. (1989). Spacing effects in memory: Evidence for a two-process account. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15*, 371–377. http://dx.doi.org/10.1037/0278-7393.15.3.371

Große, C. S., & Renkl, A. (2007). Finding and fixing errors in worked examples: Can this foster learning outcomes?. *Learning and Instruction, 17*, 612–634. https://doi.org/10.1016/j.learninstruc.2007.09.008

Halamish, V., & Bjork, R. A. (2011). When does testing enhance retention? A distribution-based interpretation of retrieval as a memory modifier. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*, 801–812. http://dx.doi.org/10.1037/a0023219

Halpern, D. F. (1998). Teaching critical thinking for transfer across domains: Disposition, skills, structure training, and metacognitive monitoring. *American Psychologist, 53*, 449–455. https://doi.org/10.1002/tl.8005

Halpern, D. F. (2014). *Critical thinking across the curriculum: A brief edition of thought & knowledge.* Routledge.

Halpern, D. F., & Butler, H. A. (2019). Teaching critical thinking as if our future depends on it, because it does. In J. Dunlosky & K. Rawson (Eds.), *The Cambridge handbook of cognition and education* (pp. 51–66). Cambridge University Press.

Haskell, R. E. (2001). *Transfer of learning: Cognition, instruction, and reasoning*. Academic Press.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*, 81–112. https://doi.org/10.3102/003465430298487

HBO-raad. (2009). *Kwaliteit als Opdracht*. The Hague: HBO-raad.

Heijltjes, A., Van Gog, T., Leppink, J., & Paas, F. (2014a). Improving critical thinking: Effects of dispositions and instructions on economics students' reasoning skills. *Learning and Instruction, 29*, 31–42. https://doi.org/10.1016/j.learninstruc.2013.07.003

Heijltjes, A., Van Gog, T., Leppink, J., & Paas, F. (2015). Unraveling the effects of critical thinking instructions, practice, and self-explanation on students' reasoning performance. *Instructional Science, 43*, 487–506. https://doi.org/10.1002/acp.3025

Heijltjes, A., Van Gog, T., & Paas, F. (2014b). Improving students' critical thinking: Empirical support for explicit instructions combined with practice. *Applied Cognitive Psychology, 28*, 518–530. https://doi.org/10.1002/acp.3025

Helsdingen, A., Van Gog, T., & Van Merriënboer, J. (2011a). The effects of practice schedule and critical thinking prompts on learning and transfer of a complex judgment task. *Journal of Educational Psychology, 103*, 383–398. https://doi.org/10.1037/a0022370

Helsdingen, A. S., Van Gog, T., & van Merriënboer, J. J. G. (2011b). The effects of practice schedule on learning a complex judgment task. *Learning and Instruction, 21*, 126–136. https://doi.org/10.1016/j.learninstruc.2009.12.001

Herzog, S. M., & Hertwig, R. (2009). The wisdom of many in one mind: Improving individual judgments with dialectical bootstrapping. *Psychological Science, 20*, 231-237. https://doi.org/10.1111%2Fj.1467-9280.2009.02271.x

Hintzman, D. L. (1976). Repetition and memory. In G. H. Bower (Ed), *Psychology of learning and motivation* (Vol. 10, pp. 47–91). Academic Press.

Hintzman, D. L. (2010). How does repetition affect memory? Evidence from judgments of recency. *Memory & Cognition, 38*, 102–115. https://doi.org/10.3758/MC.38.1.102

Hintzman, D. L., & Block, R. A. (1971). Repetition and memory: Evidence for a multiple-trace hypothesis. *Journal of Experimental Psychology, 88*, 297–306. https://doi.org/10.1037/h0030907

Hoffman, B., & Schraw, G. (2010). Conceptions of efficiency: Applications in learning and problem solving. *Educational Psychologist, 45*, 1–14. https://doi.org/10.1080/00461520903213618

Holland, J. H., Holyoak, K. J., Nisbett, R. E., & Thagard, P. R. (1989). *Induction: Processes of inference, learning, and discovery*. MIT press.

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

Huber, C. R., & Kuncel, N. R. (2016). Does college teach critical thinking? *A meta-analysis. Review of Educational Research, 86*, 431-468. https://doi.org/10.3102/0034654315605917

Hulleman, C. S., & Cordray, D. S. (2009). Moving from the lab to the field: The role of fidelity and achieved relative intervention strength. *Journal of Research on Educational Effectiveness, 2*, 88–110. https://doi.org/10.1080/19345740802539325

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

Isotani, S., Adams, D., Mayer, R. E., Durkin, K., Rittle-Johnson, B., & McLaren, B. M. (2011). Can erroneous examples help middle-school students learn decimals? *In Proceedings of the sixth European conference on technology enhanced learning: Towards ubiquitous learning (EC-TEL-2011).*

Ikuenobe, P. (2001). Teaching and Assessing Critical Thinking Abilities as Outcomes in an Informal Logic Course. *Teaching in Higher Education, 6*, 19–32. https://doi.org/10.1080/13562510020029572

Ioannidis, J. P., Munafo, M. R., Fusar-Poli, P., Nosek, B. A., & David, S. P. (2014). Publication and other reporting biases in cognitive sciences: Detection, prevalence, and prevention. *Trends in Cognitive Sciences, 18*, 235–241. https://doi.org/10.1016/j.tics.2014.02.010

Janiszewski, C., Noel, H., & Sawyer, A. G. (2003). A meta-analysis of the spacing effect in verbal learning: Implications for research on advertising repetition and consumer memory. *Journal of Consumer Research, 30*, 138–149. https://doi.org/10.1086/374692

Janssen, E. M., Mainhard, T., Buisman, R. S. M., Verkoeijen, P. P. J. L. Heijltjes, A. E. G., Van Peppen, L. M., & Van Gog, T. (2019a). Training higher education teachers' critical thinking and attitudes towards teaching It. *Contemporary Educational Psychology, 58*, 310–322. https://doi.org/10.1016/j.cedpsych.2019.03.007

Janssen, E. M., Meulendijks, W., Mainhard, T., Verkoeijen, P. P., Heijltjes, A. E., Van Peppen, L. M., & Van Gog, T. (2019b). Identifying characteristics associated with higher education teachers' Cognitive Reflection Test performance and their attitudes towards teaching critical thinking. *Teaching and Teacher Education, 84*, 139–149. https://dx.doi.org/10.1016/j.tate.2019.05.008

Janssen, E. M., Verkoeijen, P. P. J. L., Heijltjes, A. E. G., Mainhard, T., Van Peppen, L. M., & Van Gog, T. (2020). Psychometric properties of the Actively Open-minded Thinking scale. *Thinking Skills and Creativity, 36*, 100659.

Jennings, K. E., Amabile, T., & Ross, L. (1982). Informal covariation assessment: Data-based vs. theory-based judgments. In D. Kahneman, P. Slovic, & A. Tversky (Eds.) *Judgment under uncertainty: Heuristics and biases* (pp. 211–230). Cambridge University Press.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science, 23*, 524–532. https://doi.org/10.1177/0956797611430953

Kahneman D, & Frederick S. (2002). Representativeness revisited: attribute substitution in intuitive judgement. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 49–81). Cambridge University Press.

Kahneman, D., & Klein, G. (2009). Conditions for intuitive expertise: A failure to disagree. *American Psychologist, 64*, 515–526. https://doi.org/10.1037/a0016755

Kahneman, D., & Knetsch, J. (1993). Anchoring or shallow inferences: The effect of format. *Unpublished manuscript*, University of California, Berkeley.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology, 3*, 430–454. https://doi.org/10.1016/0010-0285(72)90016-3

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review, 80*, 237–251. https://doi.org/10.1037/h0034747\

Kalyuga, S. (2007). Expertise reversal effect and its implications for learner-tailored instruction. *Educational Psychology Review, 19*, 509–539. https://doi.org/10.1007/s10648-007-9054-3

Kalyuga, S. (2011). Cognitive load theory: How many types of load does it really need?. *Educational Psychology Review, 23*, 1–19. https://doi.org/10.1007/s10648-010-9150-7

Kalyuga, S., Ayres, P., Chandler, P., & Sweller, J. (2003). The expertise reversal effect. *Educational Psychologist, 38*, 23–31. https://doi.org/10.1207/ S15326985EP3801_4

Kalyuga, S., Rikers, R., & Paas, F. (2012). Educational implications of expertise reversal effects in learning and performance of complex cognitive and sensorimotor skills. *Educational Psychology Review, 24*, 313–337. https://doi.org/10.1007/s10648-012-9195-x

Kang, M. J., Hsu, M., Krajbich, I. M., Loewenstein, G., McClure, S. M., Wang, J. T., & J. F. Camerer (2009). The wick in the candle of learning: Epistemic curiosity activates reward circuitry and enhances memory. *Psychological Science, 20*, 963–973. https://doi.org/10.1111/j.1467-9280.2009.02402.x

Kanyongo, G. Y., Brook, G. P., Kyei-Blankson, L., & Gocmen, G. (2007). Reliability and statistical power: How measurement fallibility affects power and required sample sizes for several parametric and nonparametric statistics. *Journal of Modern Applied Statistical Methods, 6*, 81–90. https://doi.org/10.22237/jmasm/1177992480

Karpicke, J. D., & Roediger III, H. L. (2007). Repeated retrieval during learning is the key to long-term retention. *Journal of Memory and Language, 57*, 151-162. https://doi.org/10.1016/j.jml.2006.09.004

Kassin, S. M., Dror, I. E., & Kukucka, J. (2013). The forensic confirmation bias: Problems, perspectives, and proposed solutions. *Journal of Applied Research in Memory and Cognition, 2*, 42–52. https://doi.org/10.1016/j.jarmac.2013.01.001

Kaufmann, L., Carter, C. R., & Buhrmann, C. (2010). Debiasing the supplier selection decision: a taxonomy and conceptualization. *International Journal of Physical Distribution & Logistics Management, 40*, 792–821. http://doi.org/10.1108/09600031011093214

Kawasaki, M. (2010). Learning to solve mathematics problems: The impact of incorrect solutions in fifth grade peers' presentations. *Japanese Journal of Developmental Psychology, 21*, 12–22.

Kenyon, T., & Beaulac. (2014). Critical thinking education and debiasing. *Informal Logic, 34*, 341–363. https://doi.org/10.22329/il.v34i4.4203

Kintsch, W. (1994). Text comprehension, memory, and learning. *American Psychologist, 49*, 294–304. https://doi.org/10.1037/0003-066X.49.4.294

Klassen, R. M., & Tze, V. M. C. (2014). Teachers' self-efficacy, personality, and teaching effectiveness: A meta-analysis. *Educational Research Review, 12,* 59–76. https://doi.org/10.1016/j.edurev.2014.06.001.

Kliegl, O., Bjork, R. A., & Bäuml, K. H. T. (2019). Feedback at test can reverse the retrieval-effort effect. *Frontiers in Psychology, 10*, 1863. https://doi.org/10.3389/fpsyg.2019.01863

Koehler, D. J., Brenner, L., & Griffin, D. (2002). The calibration of expert judgment: Heuristics and biases beyond the laboratory. In T. Gilovich, D. W. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 686–715). Cambridge University Press.

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*, 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

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

Koriat, A., Sheffer, L., & Ma'ayan, H. (2002). Comparing objective and subjective learning curves: judgments of learning exhibit increased underconfidence with practice. *Journal of Experimental Psychology: General, 131*, 147–162. http://dx.doi.org/10.1037/0096-3445.131.2.147

Kornell, N., & Bjork, R. A. (2008). Learning concepts and categories: Is spacing the “enemy of induction”? *Psychological Science, 19*, 585–592. https://doi.org/10.1111/j.1467-9280.2008.02127.x

Kornell, N., Bjork, R. A., & Garcia, M. A. (2011). Why tests appear to prevent forgetting: A distribution-based bifurcation model. *Journal of Memory and Language, 65*, 85–97. https://doi.org/10.1016/j.jml.2011.04.002

Ku, K. Y. L. (2009). Assessing students’ critical thinking performance: Urging for measurements using multi-response format. *Thinking Skills and Creativity, 4*, 70–76. https://doi.org/10.1016/j.tsc.2009.02.001

Kuhn, D. (2005). *Education for thinking.* Harvard University Press.

Lai, E. R. (2011). Critical thinking: A literature review. *Pearson's Research Reports, 6*, 40–41.

Lansdale, M., & Baguley, T. (2008). Dilution as a model of long-term forgetting. *Psychological Review, 115*, 864–892. http://dx.doi.org/10.1037/a0013325

Larrick, R. (2004). Debiasing. In D. J. Koehler & N. Harvey (Eds.), *Blackwell handbook of judgment and decision making* (pp. 316–338). Blackwell Publishing Ltd.

LeBel, E. P., & Paunonen, S. V. (2011). Sexy but often unreliable: The impact of unreliability on the replicability of experimental findings with implicit measures. *Personality and Social Psychology Bulletin, 37*, 570–583. https://doi.org/10.1177%2F0146167211400619

Leppa, C. J. (1997). Standardized measures of critical thinking: Experience with the California Critical Thinking Tests. *Nurse Education, 22*, 29–33.

Lichtenstein, S., & Slovic, P. (1971). Reversals of preference between bids and choices in gambling decisions. *Journal of Experimental Psychology, 89*, 46–55. https://psycnet.apa.org/doi/10.1037/h0031207

Likourezos, V., Kalyuga, S., & Sweller, J. (2019). The variability effect: When instructional variability is advantageous. *Educational Psychology Review, 31*, 479–497. https://doi.org/10.1007/s10648-019-09462-8

Liu, O. L., Frankel, L., & Roohr, K. C. (2014). Assessing critical thinking in higher education: Current state and directions for next-generation assessment. *ETS Research Report Series, 2014*, 1–23. https://doi.org/10.1002/ets2.12009

Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological Review, 95*, 492–527. http://dx.doi.org/10.1037/0033-295X.95.4.492

Loibl, K., & Leuders, T. (2019). How to make failure productive: Fostering learning from errors through elaboration prompts. *Learning and Instruction, 62*, 1–10. https://doi.org/10.1016/j.learninstruc.2019.03.002

Loibl, K., & Leuders, T. (2018). Errors during exploration and consolidation—The effectiveness of productive failure as sequentially guided discovery learning. *Journal für Mathematik-Didaktik, 39*, 69–96. https://doi.org/10.1007/s13138-018-0130-7

Lombrozo, T. (2006). The structure and function of explanations. *Trends in Cognitive Sciences, 10*, 464–470. https://doi.org/10.1016/j.tics.2006.08.004

Loo, R., & Thorpe, K. (1999). A psychometric investigation of scores on the Watson-Glaser critical thinking appraisal new forms. *Educational and Psychological Measurement, 59*, 995–1003. https://doi.org/10.1177%2F00131649921970305

Lord, C. G., Lepper, M. R., & Preston, E. (1984). Considering the opposite: A corrective strategy for social judgment. *Journal of Personality and Social Psychology, 47*, 1231–1243. https://doi.org/10.1037/0022-3514.47.6.1231

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 internal medicine residents. *JAMA, 304*, 1198–1203. https://doi.org/10.1001/jama.2010.1276

Markovits, H., & Nantel, G. (1989). The belief-bias effect in the production and evaluation of logical conclusions. *Memory & Cognition, 17*, 11–17. https://doi.org/10.3758/BF03199552

Marsh, E. J., & Eliseev, E. D. (2019). Correcting student errors and misconceptions. In J. Dunlosky & K. Rawson (Eds.), *The Cambridge Handbook of Cognition and Education* (pp. 183–208). Cambridge University Press.

Mayer, R. E. (2002). Rote versus meaningful learning. *Theory into Practice, 41*, 226-232. https://doi.org/10.1207/s15430421tip4104_4

Mayer, R. E., & Wittrock, M. C. (1996). Problem-solving transfer. In D.C. Berlinert & R.C. Calfee (Eds.), *Handbook of Educational Psychology* (pp. 47–62). Macmillan.

Matthews, P., & Rittle-Johnson, B. (2009). In pursuit of knowledge: Comparing self-explanations, concepts, and procedures as pedagogical tools. *Journal of Experimental Child Psychology, 104,* 1–21. https://doi.org/10.1016/j.jecp.2008.08.004

McClelland, J. L., & Chappell, M. (1998). Familiarity breeds differentiation: A subjective likelihood approach to the effects of experience in recognition memory. *Psychological Review, 105*, 724–760. http://dx.doi.org/10.1037/0033-295X.105.4.734-760

McDaniel, M. A. (2007). Transfer: Rediscovering a central concept. In H. L. Roediger, Y. Dudai, & S. M. Fitzpatrick (Eds.), *Science of memory: Concepts.* Oxford University Press.

McDaniel, M. A., & Butler, A. C. (2010). A contextual framework for understanding when difficulties are desirable. In A. S. Benjamin (Ed.), *Successful remembering and successful forgetting: A festschrift in honor of Robert A. Bjork* (pp. 175–198). Psychology Press.

McDaniel, M. A., Thomas, R. C., Agarwal, P. K., McDermott, K. B., & Roediger, H. L. (2013). Quizzing in middle-school science: Successful transfer performance on classroom exams. *Applied Cognitive Psychology, 27*, 360–372. https://doi.org/10.1002/acp.2914

McDaniel, M. A., Wildman, K. M., & Anderson, J. L. (2012). Using quizzes to enhance summative-assessment performance in a web-based class: An experimental study. *Journal of Applied Research in Memory & Cognition, 1*, 18–26. https://doi.org/10.1016/j.jarmac.2011.10.001

McEldoon, K. L., Durkin, K. L., & Rittle-Johnson, B. (2013). Is self-explanation worth the time? A comparison to additional practice. *British Journal of Educational Psychology, 83*, 615–632. https://doi.org/10.1111/j.2044-8279.2012.02083.x

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1*, 30–46. https://doi.org/10.1037/1082-989X.1.1.30

McKenney, S., & Reeves, T. C. (2018). *Conducting educational design research*. Routledge.

McLaren, B. M., Adams, D. M., & Mayer, R. E. (2015). Delayed learning effects with erroneous examples: A study of learning decimals with a web-based tutor. *International Journal of Artificial Intelligence in Education, 25*, 520–542. https://doi.org/10.1007/s40593-015-0064-x

McLaren, B. M., Van Gog, T., Ganoe, C., Karabinos, M., & Yaron, D. (2016). The efficiency of worked examples compared to erroneous examples, tutored problem solving, and problem solving in computer-based learning environments. *Computers in Human Behavior, 55*, 87–99. https://doi.org/10.1016/j.chb.2015.08.038

McNeil, N. M., & Fyfe, E. R. (2012). "Concreteness fading" promotes transfer of mathematical knowledge. *Learning and Instruction, 22*, 440–448. https://doi.org/10.1016/j.learninstruc.2012.05.001

McPeck, J. E. (1990). Critical thinking and subject specificity: A reply to Ennis. *Educational Researcher, 19*, 10–12. https://doi.org/10.3102/0013189X019004010

McPeck, J. E. (1992). Thoughts on subject specificity. In S. Norris (Ed.), *The generalizability of critical thinking: Multiple perspectives on an educational ideal* (pp. 198–205). Teachers College Press.
Melton, A. W. (1967). Repetition and retrieval from memory. *Science, 158*, 532–532. https://doi.org/10.1126/science.158.3800.532-b.
Melton, A. W. (1970). The Situation with Respect to the Spacing of Repetitions and Memory. *Journal of Verbal Learning and Verbal Behavior, 9*, 596–606. https://doi.org/10.1016/S0022-5371(70)80107-4
Metcalfe, J. (2011). Desirable difficulties and studying in the region of proximal learning. In A. S. Benjamin (Ed.), *Successful remembering and successful forgetting: A festschrift in honor of Robert A. Bjork* (pp. 1259–276). Psychology Press.
Moore, T. (2004). The critical thinking debate: How general are general thinking skills?. *Higher Education Research & Development, 23*, 3–18. https://doi.org/10.1080/0729436032000168469
Moxley, S. E. (1979). Schema: The variability of practice hypothesis. *Journal of Motor Behavior, 11*, 65–70. http://dx.doi.org/10.1080/00222895.1979. 10735173
Murdock, B. B., Jr., Smith, D., & Bai, J. (2001). Judgments of frequency and recency in a distributed memory model. *Journal of Mathematical Psychology, 45*, 564–602. https://doi.org/10.1006/jmps.2000.1339
Nelson, T. O., Dunlosky, J., Graf, A., & Narens, L. (1994). Utilization of metacognitive judgments in the allocation of study during multitrial learning. *Psychological Science, 5*, 207–213. https://doi.org/10.1111/j.1467-9280.1994.tb00502.x
Newstead, S. E., Pollard, P., Evans, J. St. B. T., & Allen, J. L. (1992). The source of belief bias effects in syllogistic reasoning. *Cognition, 45*, 257–284. https://doi.org/10.1016/0010-0277(92)90019-E
Nievelstein, F., Van Gog, T., Van Dijck, G., & Boshuizen, H. P. (2013). The worked example and expertise reversal effect in less structured tasks: Learning to reason about legal cases. *Contemporary Educational Psychology, 38*, 118–125. https://doi.org/10.1007/s11251-008-9076-3
Niu, L., Behar-Horenstein, L. S., & Garvan, C. W. (2013). Do instructional interventions influence college students' critical thinking skills? A meta-analysis. *Educational Research Review, 9*, 114–128. https://doi.org/10.1016/j.edurev.2012.12.002
Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., ... & Contestabile, M. (2015). Promoting an open research culture. *Science, 348*, 1422–1425. https://doi.org/10.1126/science.aab2374
OCW (2019). *Strategische agenda hoger onderwijs en onderzoek: Houdbaar voor de toekomst.* The Hague: Ministerie van Onderwijs, Cultuur en Wetenschap.
OECD (2018). *The future of education and skills 2030, Education 2030.* Paris: The Organisation for Economic Co-operation and Development.
Onderwijsraad (2014a). *Een eigentijds curriculum.* The Hague: Onderwijsraad.
Onderwijsraad (2014b). *Meer innovatieve professionals.* The Hague: Onderwijsraad.
Onderwijsraad (2017). *Het bevorderen van gelijke kansen en sociale samenhang.* The Hague: Onderwijsraad.
Onderwijsraad (2018). *Ruim baan voor leraren: Een nieuw perspectief op het leraarschap.* The Hague: Onderwijsraad.
Open Science Collaboration. (2012). An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspectives on Psychological Science, 7*, 657–660. https://doi.org/10.1177%2F1745691612462588
Osborne, R. J., & Wittrock, M. C. (1983). Learning science: A generative process. *Science Education, 67*, 489–508. https://doi.org/10.1002/sce.3730670406
Paas, F. (1992). Training strategies for attaining transfer or problem solving skills in statistics: A cognitive-load approach. *Journal of Educational Psychology, 84*, 429–434. https://doi.org/10.1037/0022-0663.84.4.429

Paas, F., Renkl, A., & Sweller, J. (2003a). Cognitive load theory and instructional design: Recent developments. *Educational Psychologist, 38*, 1–4. https://doi.org/10.1207/S15326985EP3801_1

Paas, F., Tuovinen, J. E., Tabbers, H., & Van Gerven, P. W. (2003b). Cognitive load measurement as a means to advance cognitive load theory. *Educational Psychologist, 38*, 63–71. https://doi.org/10.1207/S15326985EP3801_8

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

Paas, F., & Van Merriënboer, J. J. G. (1993). The efficiency of instructional conditions: An approach to combine mental efforts and performance measures. *Human Factors, 35*, 737–743. https://doi.org/10.1177/001872089303500412

Paas, F. G., & Van Merriënboer, J. J. (1994). Variability of worked examples and transfer of geometrical problem-solving skills: A cognitive-load approach. *Journal of Educational Psychology, 86*, 122. https://psycnet.apa.org/doi/10.1037/0022-0663.86.1.122

Pan, S. C., Gopal, A., & Rickard, T. C. (2016). Testing with feedback yields potent, but piecewise, learning of history and biology facts. *Journal of Educational Psychology, 108*, 563–575. http://doi.org/10.1037/edu0000074

Pan, S. C. & Rickard, T. C. (2018). Transfer of test-enhanced learning: Meta-analytic review and synthesis. *Psychological Bulletin, 144*, 710–756. https://doi.org/10.1037/bul0000151

Pascarella, E. T., Blaich, C., Martin, G. L., & Hanson, J. M. (2011). How robust are the findings of Academically Adrift? *Change: The Magazine of Higher Learning, 43*, 20–24. https://doi.org/10.1080/00091383.2011.568898

Pastötter, B., & Bäuml, K.-H. T. (2016). Reversing the testing effect by feedback: Behavioral and electrophysiological evidence. *Cognitive, Affective, & Behavioral Neuroscience, 16*, 473–488. https://doi.org/10.3758/s13415-016-0407-6.

Pellegrino, J. W., & Hilton, M. L. (2012). *Education for life and work: Transferable knowledge and skills for the 21st century*. National Academies Press.

Perkins, D. N., Jay, E., & Tishman, S. (1993). Beyond abilities: A dispositional theory of thinking. *Merrill-Palmer Quarterly, 39*, 1–21.

Perkins, D. N., & Salomon, G. (1992). Transfer of learning. In T. Husen & T. N. Postelwhite (Eds.), *The international encyclopedia of education* (2nd ed., Vol. 11, pp. 6452–6457). Pergamon Press.

Pocock, S. J. (1992). When to stop a clinical trial. *British Medical Journal, 305*, 235–240. https://doi.org/10.1136/bmj.305.6847.235

Potts, R., & Shanks, D. R. (2019). The benefit of generating errors during learning: what is the locus of the effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 45*, 1023–1041. https://doi.org/10.1037/xlm0000637

Rachlinski, J. J. (2004). Heuristics, biases, and governance. In D. J. Koehler, & N. Harvey (Eds.), *Blackwell handbook of judgment and decision making,* (pp. 567–584). Blackwell Publishing Ltd.

Rau, M. A., Aleven, V., & Rummel, N. (2010). Blocked versus interleaved practice with multiple representations in an intelligent tutoring system for fractions. In V. Aleven, & K. J. Mostow (Eds.), *International conference on intelligent tutoring systems* (pp. 413–422). Springer.

Rau, M. A., Aleven, V., & Rummel, N. (2013). Interleaved practice in multi-dimensional learning tasks: Which dimension should we interleave? *Learning and Instruction, 23*, 98–114. https://doi.org/10.1016/j.learninstruc.2012.07.003

Rawson, K. A., & Dunlosky, J. (2011). Optimizing schedules of retrieval practice for durable and efficient learning: How much is enough? *Journal of Experimental Psychology: General, 140*, 283–302. http://dx.doi.org/10.1037/a0023956

Rawson, K. A., Dunlosky, J., & Sciartelli, S. M. (2013). The power of successive relearning: Improving performance on course exams and long-term retention. *Educational Psychology Review, 25*, 523–548. http://dx.doi.org/10.1007/s10648-013-9240-4

Rear, D. (2019) One size fits all? The limitations of standardized assessment in critical thinking. *Assessment & Evaluation in Higher Education, 44*, 664–675. https://doi.org/10.1080/02602938.2018.1526255

Reed, S. K. (1987). A structure-mapping model for word problems. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13*, 124–139. http://dx.doi.org/10.1037/0278-7393.13.1.124

Renkl, A. (1999). Learning mathematics from worked-out examples: Analyzing and fostering self-explanations. *European Journal of Psychology of Education, 14*, 477–488. https://doi.org/10.1007/BF03172974

Renkl, A. (2014). Toward an instructionally oriented theory of example-based learning. *Cognitive Science, 38*, 1–37. https://doi.org/10.1111/cogs.12086

Renkl, A., & Atkinson, R. K. (2010). Learning from worked-out examples and problem solving. In J. Plass, R., Moreno, & Brünken, R. (Eds.), *Cognitive load theory and research in educational psychology* (pp. 89–108). Cambridge University Press.

Renkl, A. & Eitel, A. (2019). Self-explaining: Learning about principles and their application. In J. Dunlosky & K. Rawson (Eds.), *The Cambridge handbook of cognition and education* (pp. 528–549). Cambridge University Press.

Renkl, A., Hilbert, T., & Schworm, S. (2009). Example-based learning in heuristic domains: A cognitive load theory account. *Educational Psychology Review, 21*, 67–78. https://doi.org/10.1007/s10648-008-9093-4

Rich, P. R., Van Loon, M. H., Dunlosky, J., & Zaragoza, M. S. (2017). Belief in corrective feedback for common misconceptions: Implications for knowledge revision. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 43*, 492–501. https://doi.org/10.1037/xlm0000322

Richland, L. E., Bjork, R. A., Finley, J. R., & Linn, M. C. (2005). Linking cognitive science to education: Generation and interleaving effects. In *Proceedings of the twenty-seventh annual conference of the Cognitive Science Society* (pp. 1850–1855). Erlbaum.

Rickard, T. C., & Pan, S. C. (2017). A dual memory theory of the testing effect. *Psychonomic Bulletin & Review, 25*, 847–869. https://doi.org/10.3758/s13423-017-1298-4

Ritchhart, R., & Perkins, D. N. (2005). Learning to think: The challenges of teaching thinking. In Holyoak, K. J., & Morrison, R. G. (Eds.), *The Cambridge handbook of thinking and reasoning* (pp. 775–802). Cambridge University Press.

Rittle-Johnson, B., & Loehr, A. (2017). Eliciting explanations: Constraints on when self-explanation aids learning. *Psychonomic Bulletin & Review, 24,* 1501–1510. https://doi.org/s13423-016-1079-5

Rittle-Johnson, B., Star, J. R., & Durkin, K. (2009). The importance of prior knowledge when comparing examples: Influences on conceptual and procedural knowledge of equation solving. *Journal of Educational Psychology, 101*, 836–852. https://psycnet.apa.org/doi/10.1037/a0016026

Robinson, S. R. (2011). Teaching logic and teaching critical thinking: revisiting McPeck. *Higher Education Research & Development, 30*, 275–287. http://doi.org/10.1080/07294360.2010.500656

Roediger, H. L., III, & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences, 15*, 20–27. https://doi.org/10.1016/j.tics.2010.09.003

Roediger, H. L., III, & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science, 1*, 181–210. https://doi.org/10.1111/j.1745-6916.2006.00012.x

Roelle, J. & Berthold, K. (2015). Effects of comparing contrasting cases on learning from subsequent explanations. *Cognition and Instruction, 33*, 199–225. https://doi.org/10.1080/07370008.2015.1063636

Rogers, W. T., & Hopkins, K. D. (1988). Power estimates in the presence of a covariate and measurement error. *Educational and Psychological Measurement, 48*, 647–656. https://doi.org/10.1177/0013164488483008

Rohrer, D., Dedrick, R. F., & Burgess, K. (2014). The benefit of interleaved mathematics practice is not limited to superficially similar kinds of problems. *Psychonomic Bulletin & Review, 21*, 1323–1330. https://doi.org/10.3758/s13423-014-0588-3

Rohrer, D., Dedrick, R. F., Hartwig, M. K., & Cheung, C.-N. (2019). A randomized controlled trial of interleaved mathematics practice. *Journal of Educational Psychology, 112*, 40–52. https://doi.org/10.1037/edu0000367

Rohrer, D., Dedrick, R. F., & Stershic, S. (2015). Interleaved practice improves mathematics learning. *Journal of Educational Psychology, 107*, 900–908. https://doi.org/10.1037/edu0000001

Rohrer, D., Taylor, K., & Sholar, B. (2010). Tests enhance the transfer of learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*, 233–239. http://dx.doi.org/10.1037/a0017678

Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin, 140*, 1432–1463. http://dx.doi.org/10.1037/a0037559

Royalty, J. (1995). The generalizability of critical thinking: Paranormal beliefs versus statistical reasoning. *The Journal of Genetic Psychology, 156*, 477–488. https://doi.org/10.1080/00221325.1995.9914838

Salomon, G., & Perkins, D. N. (1989). Rocky roads to transfer: Rethinking mechanism of a neglected phenomenon. *Educational Psychologist, 24*, 113–142. https://doi.org/10.1207/s15326985ep2402_1

Sana, F., Yan, V. X., Kim, J. A., Bjork, E. L., & Bjork, R. A. (2018). Does working memory capacity moderate the interleaving benefit?. *Journal of Applied Research in Memory and Cognition, 7*, 361–369. https://doi.org/10.1016/j.jarmac.2018.05.005

Schmidt, F.L., & Hunter, J.E. (1996). Measurement error in psychological research: Lessons from 27 research scenarios. *Psychological Methods, 1*, 199–223. https://doi.org/10.1037/1082-989X.1.2.199

Schmidt, H. G., Mamede, S., Van Den Berge, K., Van Gog, T., Van Saase, J. L., & Rikers, R. M. (2014). Exposure to media information about a disease can cause doctors to misdiagnose similar-looking clinical cases. *Academic Medicine, 89*, 285–291. https://doi.org/10.1097/ACM.0000000000000107

Schneider, V. I., Healy, A. F., & Bourne, L. E. (2002). What is learned under difficult conditions is hard to forget: Contextual interference effects in foreign vocabulary acquisition, retention, and transfer. *Journal of Memory and Language, 46*, 419–440. https://doi.org/10.1006/jmla.2001.2813

Schneider, V. I., Healy, A. F., Ericsson, K. A., & Bourne Jr, L. E. (1995). The effects of contextual interference on the acquisition and retention of logical rules. In A. F. Healy & L. E. Bourne, Jr. (Eds.), *Learning and memory of knowledge and skills: Durability and specificity* (pp. 95–131). Sage Publications, Inc.

Schworm, S., & Renkl, A. (2007). Learning argumentation skills through the use of prompts for self-explaining examples. *Journal of Educational Psychology, 99*, 285–296. https://psycnet.apa.org/doi/10.1037/0022-0663.99.2.285

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory. *Psychological Review, 84*, 127–190. https://psycnet.apa.org/doi/10.1037/0033-295X.84.2.127

Siegler, R.S. (2002). Microgenetic studies of self-explanations. In N. Grannot & J. Parziale (Eds.), *Microdevelopment: Transition processs in development and learning* (pp. 31–58). Cambridge University Press.

Simmons, J.P., Nelson, L.D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*, 1359–1366. https://doi.org/10.1177/0956797611417632

Slamecka, N. J., & Graf, P. (1978). The generation effect: Delineation of phenomenon. *Journal of Experimental Psychology: Human Learning and Memory, 4,* 592–604.

Slovic, P. (1967). The relative influence of probabilities and payoffs upon perceived risk of a gamble. *Psychonomic Science, 9*, 223–224. https://doi.org/10.3758/BF03330840

Smith, G. (2002). Are there domain-specific thinking skills? *Journal of Philosophy of Education, 36*, 207–227. http://doi.org/doi:10.1111/1467-9752.00270

Soderstrom, N. C., & Bjork, R. A. (2015). Learning versus performance: An integrative review. *Perspectives on Psychological Science, 10*, 176–199. https://doi.org/10.1177/1745691615569000

Stanovich, K. E. (1999). *Who is rational? Studies of individual differences in reasoning.* Psychology Press.

Stanovich, K. E. (2005). *The robot's rebellion: Finding meaning in the age of Darwin.* University of Chicago Press.

Stanovich, K. E. (2011). *Rationality and the reflective mind.* Oxford University Press.

Stanovich, K. E., Toplak, M. E., & West, R. F. (2008). The development of rational thought: A taxonomy of heuristics and biases. *Advances in Child Development and Behavior, 36*, 251–285. https://doi.org/10.1016/S0065-2407(08)00006-2

Stanovich, K. E. & West, R. F. (1999). Discrepancies between normative and descriptive models of decision making and the understanding/acceptance principle. *Cognitive Psychology, 38*, 349–385. https://doi.org/10.1006/cogp.1998.0700

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate?. *Behavioral and Brain Sciences, 23*, 645–665. https://doi.org/10.1017/S0140525X00003435

Stanovich, K. E., & West, R. F. (2007). Natural myside bias is independent of cognitive ability. *Thinking & Reasoning, 13*, 225–247. https://doi.org/10.1080/13546780600780796

Stanovich, K. E., West, R. K., & Toplak, M. E. (2016). *The rationality quotient: Toward a test of rational thinking*. MIT Press.

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*, 22–33. https://doi.org/10.1016/j.learninstruc.2009.10.001.

Sternberg, R. J. (2001). Why schools should teach for wisdom: The balance theory of wisdom in educational settings. *Educational Psychologist, 36*, 227–245. https://doi.org/10.1207/S15326985EP3604_2

Storm, B. C., Friedman, M. C., Murayama, K., & Bjork, R. A. (2014). On the transfer of prior tests or study events to subsequent study. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*, 115–124. http://dx.doi.org/10.1037/a0034252

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science, 12*, 257–285. https://doi.org/10.1207/s15516709cog1202_4

Sweller, J., Ayres, P., & Kalyuga, S. (2011). Measuring cognitive load. In *Cognitive load theory. Explorations in the learning sciences, Instructional systems and performance technologies* (pp. 71–85). Springer.

Sweller, J., Van Merriënboer, J. J., & Paas, F. G. (1998). Cognitive architecture and instructional design. *Educational Psychology Review, 10*, 251–296.

Thiede, K. W., Anderson, M. C. M., & Therriault, D. (2003). Accuracy of metacognitive monitoring affects learning of texts. *Journal of Educational Psychology, 95*, 66–73. http://dx.doi.org/10.1037/0022-0663.95.1.66

Tiruneh, D. T., Verburgh, A., & Elen, J. (2014). Effectiveness of critical thinking instruction in higher education: A systematic review of intervention studies. *Higher Education Studies, 4*, 1–17. http://dx.doi.org/10.5539/hes.v4n1p1

Tiruneh, D. T., Weldeslassie, A. G., Kassa, A., Tefera, Z., De Cock, M., & Elen, J. (2016). Systematic design of a learning environment for domain-specific and domain-general critical thinking skills. *Educational Technology Research and Development, 64*, 481–505. https://doi.org/10.1007/s11423-015-9417-2

Tishman, S., Jay, E., & Perkins, D. N. (1993). Teaching thinking dispositions: From transmission to enculturation. *Theory Into Practice, 32*, 147–153. https://doi.org/10.1080/00405849309543590

Toppino, T. C., & Bloom, L. C. (2002). The spacing effect, free recall, and two-process theory: A closer look. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*, 437–444. https://doi.org/10.1037/0278-7393.28.3.437

Trauzettel-Klosinski, S., & Dietz, K. (2012). Standardized assessment of reading performance: the new International Reading Speed Texts IReST. *Investigative Ophthalmology & Visual Science, 53*, 5452–5461. https://doi.org/10.1167/iovs.11-8284

Tsui, L. (2002). Fostering critical thinking through effective pedagogy: Evidence from four institutional case studies. *The Journal of Higher Education, 73*, 740–763. http://dx.doi.org/10.1353/jhe.2002.0056

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology, 5*, 207–232. https://doi.org/10.1016/0010-0285(73)90033-9

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*, 1124–1131. https://doi.org/10.1126/science.185.4157.1124

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review, 90*, 293–315. https://psycnet.apa.org/doi/10.1037/0033-295X.90.4.293

Van Brussel, S., Timmermans, M., Verkoeijen, P., & Paas, F. (2020). 'Consider the opposite'–Effects of elaborative feedback and correct answer feedback on reducing confirmation bias–A pre-registered study. *Contemporary Educational Psychology, 101844*. https://doi.org/10.1016/j.cedpsych.2020.101844

Van den Broek, P., & Kendeou, P. (2008). Cognitive processes in comprehension of science texts: The role of co-activation in confronting misconceptions. *Applied Cognitive Psychology, 22*, 335–351. https://doi.org/10.1002/acp.1418

Van Gelder, T. V. (2005). Teaching critical thinking: Some lessons from cognitive science. *College Teaching, 53*, 41–48. https://doi.org/10.3200/CTCH.53.1.41-48

Van Gog, T., & Paas, F. (2008). Instructional efficiency: Revisiting the original construct in educational research. *Educational Psychologist, 43*, 16–26. https://doi.org/10.1080/00461520701756248

Van Gog, T., Paas, F., & Van Merriënboer, J. J. (2004). Process-oriented worked examples: Improving transfer performance through enhanced understanding. *Instructional Science, 32*, 83–98. https://doi.org/10.1023/B:TRUC.0000021810.70784.b0

Van Gog, T., & Rummel, N. (2010). Example-based learning: Integrating cognitive and social-cognitive research perspectives. *Educational Psychology Review, 22*, 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. Rawson (Eds.), *The Cambridge handbook of cognition and education* (pp. 183–208). Cambridge University Press.

Van Gog, T., & Sweller, J. (2015). Not new, but nearly forgotten: The testing effect decreases or even disappears as the complexity of learning materials increases. *Educational Psychology Review, 27*, 247–264. https://doi.org/10.1007/s10648-015-9310-x

VanLehn, K. (1999). Rule-learning events in the acquisition of a complex skill: An evaluation of cascade. *The Journal of the Learning Sciences, 8*, 71–125. https://doi.org/10.1207/s15327809jls0801_3.
Van Loon-Hillen, N. H., Van Gog, T., & Brand-Gruwel, S. (2012). Effects of worked examples in a primary school mathematics curriculum. *Interactive Learning Environments, 20*, 89–99. https://doi.org/10.1080/10494821003755510
Van Merriënboer, J. J. G., De Croock, M. B., & Jelsma, O. (1997). The transfer paradox: Effects of contextual interference on retention and transfer performance of a complex cognitive skill. *Perceptual and Motor Skills, 84*, 784–786. https://doi.org/10.2466/pms.1997.84.3.784
Van Merriënboer, J. J. G., Jelsma, O., & Paas, F. G. W. C. (1992). Training for reflective expertise: A four-component instructional design model for complex cognitive skills. *Educational Technology Research and Development, 40*, 23-43. https://doi.org/10.1007/bf02297047
Van Merriënboer, J. J. G., Schuurman, J. G., de Croock, M. B., & Paas, F. (2002). Redirecting learners' attention during training: Effects on cognitive load, transfer test performance and training efficiency. *Learning and Instruction, 12*, 11–37. https://doi.org/10.1016/S0959-4752(01)00020-2
Van Peppen, L. M., Verkoeijen P. P. J. L., Heijltjes, A. E. G., Janssen, E. M., Koopmans, D., & Van Gog, T. (2018). Effects of self-explaining on learning and transfer of critical thinking skills. *Frontiers in Education, 3*, 100. https://doi.org/10.3389/feduc.2018.00100.
Verburgh, A. (2013). *Research integration in higher education: Prevalence and relationship with critical thinking* (Doctoral dissertation). KU Leuven: Leuven.
Vereniging Hogescholen (2015). *Strategische visie '#hbo2025: wendbaar & weerbaar'*. The Hague: Vereniging Hogescholen.
Verkoeijen, P. P., Rikers, R. M., & Schmidt, H. G. (2004). Detrimental influence of contextual change on spacing effects in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*, 796–800. http://dx.doi.org/10.1037/0278-7393.30.4.796
Verkoeijen, P. P., Rikers, R. M., & Schmidt, H. G. (2005). Limitations to the spacing effect: Demonstration of an inverted u-shaped relationship between interrepetition spacing and free recall. *Experimental Psychology, 52*, 257–263. https://doi.org/10.1027/1618-3169.52.4.257
Vincent-Lancrin, S., González-Sancho, C., Bouckaert, M., De Luca, F., Fernández-Barrerra, M.., Gwénaël J., Urger, J., & Vidal, Q. (2019). *Fostering students' creativity and critical thinking: What it means in schools.* Educational research and innovation, OECD Publishing. https://doi.org/10.1787/62212c37-en.
Vosniadou, S., & Ortony, A. (1989). *Similarity and analogical reasoning*. Cambridge University Press.
Wahlheim, C. N., Dunlosky, J., & Jacoby, L. L. (2011). Spacing enhances the learning of natural concepts: An investigation of mechanisms, metacognition, and aging. *Memory & Cognition, 39*, 750–763. https://doi.org/10.3758/s13421-010-0063-y
Wasserman, E. A., Dorner, W. W., & Kao, S. F. (1990). Contributions of specific cell information to judgments of interevent contingency. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*, 509–521. https://doi.org/10.1037/0278-7393.16.3.509
West, R. F., Toplak, M. E., & Stanovich, K. E. (2008). Heuristics and biases as measures of critical thinking: Associations with cognitive ability and thinking dispositions. *Journal of Educational Psychology, 100*, 930–941. https://doi.org/10.1037/ a0012842
Weissgerber, S. C., Reinhard, M. A., & Schindler, S. (2018). Learning the hard way: Need for Cognition influences attitudes toward and self-reported use of desirable difficulties. *Educational Psychology, 38*, 176–202. https://doi.org/10.1080/01443410.2017.1387644
Wittrock, M. C. (1974). Learning as a generative process. *Educational Psychologist, 11*, 87–95. https://doi.org/10.1080/00461527409529129
Wittrock, M. C. (1990). Generative processes of comprehension. *Educational Psychologist, 2*4, 345–376. https://doi.org/10.1207/s15326985ep2404_2

Wittrock, M. C. (1992). Generative learning processes of the brain. *Educational Psychologist, 27*, 531–541. https://doi.org/10.1207/s15326985ep2704_8

Wittrock, M. C. (2010). Learning as a generative process. *Educational Psychologist, 45*, 40–45. http://dx.doi.org/10.1080/00461520903433554

Wylie, R., & Chi, M. T. H. (2014). The self-explanation principle in multimedia learning. In R. E. Mayer (Ed.), *Cambridge handbook of multimedia learning* (pp. 413–432). Cambridge University Press.

Ximénez, C., & Revuelta, J. (2007). Extending the CLAST sequential rule to one-way ANOVA under group sampling. *Behavior Research Methods, Instruments, & Computers, 39*, 86–100. https://doi.org/10.3758/BF03192847

Yan, V. X., Clark, C. M., & Bjork, R. A. (2016). Memory and metamemory considerations in the instruction of human beings revisited: Implications for optimizing online learning. In J. C. Horvath, J. Lodge, & J. A. C. Hattie (Eds.), *From the laboratory to the classroom: Translating the learning sciences for teachers*. Routledge.

Zelazo, P. D. (2004). The development of conscious control in childhood. *Trends in Cognitive Sciences, 8*, 12–17. http://doi.org/10.1016/j.tics.2003.11.001

Zimmerman, B. J. (2000). Self-efficacy: An essential motive to learn. *Contemporary Educational Psychology, 25*, 82–91. https://doi.org/10.1006/ceps.1999.1016

Samenvatting

Summary in Dutch

Elke dag nemen we allerlei beslissingen en vellen we oordelen. Wanneer je de trein in stapt, maak je bijvoorbeeld een besluit waar je gaat zitten. En hoor je in die trein iemand veelvoudig niezen, dan ga je er momenteel al snel van uit dat diegene een virus opgelopen heeft. Doordat we vaak beperkt zijn in de tijd en in de hoeveelheid informatie die we tot onze beschikking hebben, maken we in ons denken gebruik van *heuristieken* (ofwel vuistregels). Heuristieken helpen ons om de grote hoeveelheid informatie die we dagelijks tegenkomen aan te kunnen en om redeneerprocessen te vereenvoudigen. Daardoor kunnen we relatief snel beslissingen nemen, vaak zonder dat we ons er bewust van zijn. Maar heuristieken maken ons ook vatbaar voor systematische redeneerfouten, die ook wel *biases* worden genoemd (Tversky & Kahneman, 1974). Wanneer men bijvoorbeeld gevraagd wordt om in te schatten of het waarschijnlijker is dat iemand overlijdt aan het coronavirus of dat iemand overlijdt aan het coronavirus én ouder is dan 70 jaar, dan zal de intuïtieve reactie van de meeste mensen zijn dat de tweede optie waarschijnlijker is. Doordat de relatie tussen overlijden aan het coronavirus en het hebben van een hogere leeftijd vaak genoemd wordt en dus herkenbaar is, hebben we de neiging om de kans op deze combinatie te overschatten: we maken in dit geval gebruik van de *representativiteitsheuristiek* (Tverksy & Kahneman, 1983). De kans dat een bepaalde combinatie van gebeurtenissen voorkomt is echter áltijd kleiner dan de kans dat slechts een van deze gebeurtenissen voorkomt. Het is dus waarschijnlijker dat iemand overlijdt aan het coronavirus, dan dat diegene ook nog ouder is dan 70 jaar. In dit geval leidt het gebruik van een heuristiek dus tot een systematische redeneerfout.

Het gebruik van heuristieken kan leiden tot redeneerfouten met ernstige gevolgen. Zeker in de complexe beroepssituaties waarin de meeste afgestudeerden in het hoger onderwijs terecht komen, zoals in de medische, economische of juridische sector. Denk bijvoorbeeld aan het verkeerd toedienen van medicatie, het geven van onjuist financieel advies of het onterecht veroordelen van een verdachte voor een strafbaar feit. Het tegengaan van systematische redeneerfouten vereist dat een intuïtieve reactie wordt onderdrukt en wordt vervangen door een rationele reactie, die gebaseerd is op redeneerregels of -strategieën (uit de logica en waarschijnlijkheidstheorie; Kahneman & Tversky, 1972, 1973; Tversky & Kahneman, 1974). Ofwel, dat je kritisch kunt denken. Kritisch denken betekent, kort gezegd, dat je "redeneert en reflecteert voordat je een standpunt inneemt of een besluit neemt hoe te handelen en dat je kunt verklaren waarop dat standpunt of het besluit is gebaseerd"[15]. Het is dus van belang dat studenten in het hoger onderwijs worden opgeleid tot kritisch denkende professionals (Davies, 2013;

[15] Deze werkdefinitie is gebaseerd op de toonaangevende definitie van kritisch denken voor onderwijs en onderzoek, die is opgesteld door een panel van deskundigen: "kritisch denken wordt beschouwd als het vermogen om doelgericht, zelfregulerend te oordelen, resulterend in interpretatie, analyse, evaluatie en gevolgtrekking, alsook het verklaren waarop dat oordeel is gebaseerd in termen van bewijzen, concepten, methodes, criteria en contextuele overwegingen" (APA: Facione, 1990, p.2).

Facione, 1990; Halpern, 2014; Van Gelder, 2005). Een belangrijk kenmerk van een kritisch denkende professional is dat hij/zij in staat is om onbevooroordeeld te redeneren en beslissingen te nemen, zonder systematische redeneerfouten te maken. Er bestaan echter tal van systematische redeneerfouten en het is niet haalbaar om studenten te trainen in het vermijden van elk type redeneerfout. Daarom is het de uitdaging leeractiviteiten zó te ontwerpen dat de kritisch-denken-vaardigheden van studenten niet alleen verbeteren op de getrainde redeneertaken in een gegeven context, maar dat de geleerde vaardigheden ook tot verbetering leiden op andersoortige redeneertaken of tot een verbetering van de getrainde redeneertaken in een andere context. Met andere woorden, het doel van kritisch-denken-instructie is onder meer dat er *transfer* optreedt van de geleerde vaardigheden naar nieuwe taken en situaties. De vraag die in dit proefschrift centraal stond, was dan ook hoe kritisch-denken-instructie het beste kan worden vormgegeven om ervoor te zorgen dat studenten in het hoger onderwijs (1) leren om systematische redeneerfouten te vermijden en (2) het geleerde kunnen toepassen op nieuwe redeneertaken en in nieuwe situaties (transfer).

Om nieuwe leerstof langere tijd te onthouden en te kunnen toepassen in nieuwe situaties, moet de leerstof actief verwerkt worden. Zogenoemde 'generatieve verwerkingsstrategieën' (Engels: *generative processing strategies*) kunnen hieraan bijdragen: ze vereisen van studenten dat zij extra inspanningen leveren tijdens het leren (bijvoorbeeld door het genereren van verklaringen of vergelijkingen) en zorgen ervoor dat betekenis wordt gegeven aan de leerstof. Generatieve verwerking helpt om informatie in het geheugen te organiseren in samenhangende kennisstructuren en te integreren met reeds aanwezige kennis (Grabowski, 1996; Osborne & Wittrock, 1983; Wittrock, 1974, 1990, 1992, 2010). Bovendien kan het studenten helpen om de onderliggende principes van een probleem te identificeren en te leren welke oplossingsprocedure voor dit type probleem nodig is. Als een nieuw probleem vervolgens hetzelfde onderliggende principe heeft en de student herkent dit, dan kan hij/zij de aangeleerde procedure gebruiken om het nieuwe probleem op te lossen. Er treedt dan transfer op. Generatieve verwerkingsstrategieën zijn effectief gebleken voor het leren en de transfer van diverse vaardigheden (zie bijv. Fiorella & Mayer, 2016; Wittrock, 2010), maar het was nog onduidelijk of ze ook helpen bij het leren om systematische redeneerfouten te vermijden. In de studies in hoofdstuk 2 tot en met 5 werd daarom onderzocht of generatieve verwerkingsstrategieën eveneens het leren en de transfer van kritisch-denken-vaardigheden verder verbeteren (bovenop de effecten van instructie en oefening). In de studie in hoofdstuk 6 is daarnaast onderzocht welke factoren succesvolle transfer van kritisch-denken-vaardigheden belemmeren.

De generatieve verwerkingsstrategieën die in dit proefschrift zijn onderzocht, waren: studenten aansporen om aan zichzelf hun redeneerproces uit te leggen tijdens het oefenen, ook wel ‘zelfverklaren’ genoemd (hoofdstuk 2); variatie aanbrengen in taaktypen tijdens oefening, waarmee vergelijkingen tussen taken maken (impliciet) wordt aangemoedigd (hoofdstuk 3); studenten stimuleren om correcte en incorrecte ‘uitgewerkte voorbeelden’ te vergelijken (hoofdstuk 4); en op meerdere momenten oefentaken aanbieden aan studenten, zodat zij die informatie herhaaldelijk uit hun geheugen moeten ophalen (hoofdstuk 5). De effecten van deze strategieën, ofwel interventies, werden getest in experimenten die plaatsvonden in de praktijk van het hoger onderwijs. Zo werd in de studie in hoofdstuk 2 één groep studenten aangezet tot zelfverklaren. In elke studie was er minstens één controlegroep, die een andere of geen interventie kreeg. Eerst werd bij alle groepen een voormeting (pretest) afgenomen. Vervolgens kregen ze instructies over kritisch denken (het belang en de kenmerken van kritisch denken en de vaardigheden en houding die nodig zijn om kritisch te denken) en over specifieke heuristics-and-biases taken, vergelijkbaar met het ‘coronavirusvoorbeeld’ aan het begin van dit hoofdstuk. Daarna werd er geoefend, al dan niet met extra interventie. Direct na het oefenen werd er een nameting (posttest) afgenomen. In elk hoofdstuk werd dus op tenminste twee momenten de mate van onbevooroordeeld redeneren en de mentale inspanning gemeten (pretest en posttest). Onbevooroordeeld redeneren werd in kaart gebracht door prestatie op heuristics-and-biases taken van de test te bepalen, zowel voor taakcategorieën die deel uitmaakten van de oefenfase (hiermee werd het *leren* gemeten) als voor nieuwe taakcategorieën met dezelfde onderliggende principes (hiermee werd *transfer* gemeten). In hoofdstuk 2 en 4 werden studenten tevens op een later moment, tussen de twee weken en negen maanden, getest (verlate posttest).

De hoofdbevindingen

In hoofdstuk 2 werd onderzocht of het aanzetten van studenten tot zelfverklaren, ofwel het aan zichzelf uitleggen van hun redeneerproces (Bisra et al., 2018; Chi, 2000; Fiorella & Mayer, 2016) tijdens het oefenen, effectief zou zijn voor het leren en de transfer van de vaardigheid om systematische redeneerfouten te vermijden. De studie werd uitgevoerd in de context van een hbo-vak. De studenten ontvingen eerst de instructies over kritisch denken. Vervolgens oefenden ze met een aantal taken in de context van domeinrelevante problemen. Dat wil zeggen dat de taken realistische problemen bevatten uit het studiedomein van de studenten, in dit geval Integrale Veiligheidskunde. Tijdens het oefenen werd de helft van de studenten gevraagd om zelfverklaringen te genereren. Uit de resultaten bleek dat de prestaties van studenten op de leertaken verbeterden van pretest naar posttest en dat dit prestatieniveau na twee weken nog even hoog was. Dat wil zeggen dat studenten na de instructie en het oefenen beter in staat waren om systematische redeneerfouten te vermijden op de leertaken. Er was echter geen verschil in prestaties op de leer- en transfertaken tussen de groep die werd aangezet tot het genereren van zelfverklaringen en de controlegroep die niet werd aangezet tot zelfverklaren. Bovendien was het opmerkelijk dat de mentale inspanning die studenten leverden tijdens het maken van de taken niet verschilde tussen de twee groepen. Mogelijk hebben de studenten in de controlegroep ook generatieve verwerkingsprocessen gebruikt, bijvoorbeeld door spontaan zelfverklaringen te genereren. In dit hoofdstuk is daarnaast onderzocht of de kwaliteit van de zelfverklaringen van de studenten gerelateerd was aan hun prestaties op de leer- en transfertaken. Dit was inderdaad het geval: de studenten die verklaringen van hogere kwaliteit gaven, presteerden ook beter op de leertaken op de posttest (maar niet op de transfertaken). Eenvoudiger gezegd, de studenten die beter waren in het aan zichzelf uitleggen van hun redeneerproces, presteerden ook beter. Deze correlatie duidt wellicht op een causaal verband: door aan zichzelf hun eigen redeneerproces uit te leggen, gaan studenten beter presteren. Het kan echter ook zo zijn dat studenten met een hoger algemeen kennis- of vaardigheidsniveau betere zelfverklaringen genereren en beter presteren dan mensen met een lager kennisniveau.

In hoofdstuk 3 werd onderzocht of het creëren van variatie in oefening effectief zou zijn voor het bevorderen van leren en transfer. Er werden twee experimenten uitgevoerd; een met universitaire studenten in het laboratorium en een met hbo-studenten in de context van een vak. De variatie in oefening werd gecreëerd door de oefentaken af te wisselen die betrekking hadden op verschillende redeneerfouten (Engels: *interleaved practice*; bijv. Barreiros et al., 2007; Helsdingen et al., 2011; Rau et al., 2013) in plaats van de oefentaken gegroepeerd per redeneerfout aan te bieden (Engels: *blocked practice*). Bij een gevarieerd oefenschema werden de verschillende type oefentaken dus afgewisseld – ABACBCAABC – terwijl bij een gegroepeerd oefenschema blokken met dezelfde type oefentaken werden aangeboden – AAA-BBB-CCC. Afwisseling in taaktypen is belangrijk om studenten te leren verschillende oplossingsprocedures te gebruiken: bij elke taak moet immers het type probleem en een passende oplossing herkend worden. Hoewel dit bijdraagt aan betere prestaties op de langere termijn (bijv. Helsdingen et al., 2011a, 2011b), doet het een groter beroep op het werkgeheugen dan oefenen in een gegroepeerd schema. Omdat een te hoge werkgeheugenbelasting het leren kan belemmeren (Paas et al., 2003a), is tevens onderzocht of studenten meer zouden profiteren van afwisseling in taaktypen, wanneer ze uitgewerkte voorbeelden bestudeerden in plaats van dat ze oefenproblemen oplosten (vgl. Paas & Van Merriënboer, 1994). Het bestuderen van uitgewerkte voorbeelden – dit zijn oefeningen waarvan de oplossing volledig is uitgeschreven – leidt namelijk tot een lagere belasting van het werkgeheugen, terwijl de leerprestaties gelijk blijven of zelfs verbeteren (Van Gog et al., 2019). Nadat de studenten de instructie over kritisch denken ontvingen, volgden zij of (1) een gevarieerd oefenschema met uitgewerkte voorbeelden, of (2) een gevarieerd oefenschema met probleem-oplostaken, of (3) een gegroepeerd oefenschema met uitgewerkte voorbeelden of (4) een gegroepeerd oefenschema met probleem-oplostaken. In beide experimenten verbeterden de prestaties op de leertaken opnieuw na de instructie en het oefenen. Er waren echter geen aanwijzingen dat een gevarieerd oefenschema tot betere prestaties leidde op de leer- of transfertaken dan een gegroepeerd oefenschema, ongeacht of er geoefend werd met uitgewerkte voorbeelden of probleem-oplostaken. Een interessante bevinding uit het experiment met de universitaire studenten was dat het bestuderen van uitgewerkte voorbeelden tot betere prestaties op de leertaken leidde dan het oplossen van oefenproblemen. Bovendien werden deze prestaties bereikt met minder mentale inspanning tijdens de tests; dat wil zeggen dat de uitgewerkte voorbeelden zowel effectiever als efficiënter waren (Hoffman & Schraw, 2010; Paas & Van Merriënboer, 1993; Van Gog & Paas, 2008). Het tweede experiment met de hbo-studenten repliceerde dit positieve effect van uitgewerkte voorbeelden maar alleen bij beginners (studenten die nog weinig of geen voorkennis hadden) en niet bij meer gevorderden. Deze experimenten toonden daarmee voor het eerst aan dat het effect van uitgewerkte voorbeelden ook van toepassing is op het trainen van kritisch-denken-vaardigheden van beginners. De bevinding uit het tweede experiment is bovendien in lijn met het *expertise reversal effect* (bijv. Kalyuga et al., 2003, 2012), dat stelt dat instructiestrategieën die studenten helpen bij het ontwikkelen van cognitieve schema's effectief zijn wanneer studenten hun eerste stappen zetten in het verwerven van nieuwe kennis of vaardigheden, maar vaak niet als zij al meer gevorderd zijn.

In hoofdstuk 4 stond het vergelijken van correcte en incorrecte uitgewerkte voorbeelden (ook wel contrasterende voorbeelden genoemd) centraal. Er werd onderzocht of (1) het bestuderen van contrasterende voorbeelden zou zorgen voor een grotere verbetering in de vaardigheid om systematische redeneerfouten te vermijden dan (2) het alleen bestuderen van correcte voorbeelden, of (3) het alleen bestuderen van incorrecte voorbeelden of (4) het oplossen van oefenproblemen. De studie werd wederom uitgevoerd in de context van een hbo-vak. De studenten ontvingen eerst de instructies over kritisch denken. Daarna oefenden ze met een aantal taken in de context van domeinrelevante problemen, onder een van de vier bovengenoemde oefencondities. Uit de resultaten bleek dat de prestaties van de studenten op de leertaken wederom verbeterden na de instructie en het oefenen. Bovendien presteerden de studenten zelfs nog beter na drie weken. Er waren echter geen verschillen tussen de vier oefengroepen in prestaties op de leer-en transfertaken. Bovendien waren er in deze studie geen aanwijzingen dat het bestuderen van uitgewerkte voorbeelden tot betere prestaties leidde dan het oplossen van oefenproblemen. Dit in tegenstelling tot de bevinding uit hoofdstuk 3 en de bevindingen uit eerdere studies met vele andere soorten taken (bijv. Renkl, 2014; Van Gog et al., 2019).

De studie in hoofdstuk 5 onderzocht of het herhaaldelijk ophalen van informatie uit het geheugen (Engels: *repeated retrieval practice*) effectief zou zijn voor het leren vermijden van systematische redeneerfouten en of dit bijdraagt aan transfer. Meer specifiek, kregen studenten de mogelijkheid om te oefenen in meerdere sessies die verspreid waren over een periode van een aantal weken. Deze studie werd wederom uitgevoerd in de context van een hbo-vak. De studenten ontvingen eerst de instructies over kritisch denken. Daarna oefenden zij met een aantal taken in de context van domeinrelevante problemen. Na elke oefening werd getoond of het gegeven antwoord juist was en kregen de studenten een uitgewerkt voorbeeld van een goede redenering te zien (feedback). Afhankelijk van de groep waarin de studenten ingedeeld waren, oefenden ze eenmalig, oefenden ze een tweede keer (een week later) of oefenden ze een derde keer (twee weken na de tweede keer). Een verrassende bevinding was dat (herhaald) oefenen geen significant effect had op de prestatie op de leertaken: de drie groepen lieten geen vooruitgang zien na de instructie en het oefenen. Tenminste, dit was het geval wanneer de prestatie werd gemeten aan de hand van zowel de antwoorden op meerkeuzevragen als de onderbouwing van deze antwoorden. Wanneer alleen naar de antwoorden op de meerkeuzevragen werd gekeken, werd er wel bij alle drie de groepen een vooruitgang in prestatie op de leertaken gevonden. Bovendien bleek, zoals werd verwacht op basis van bevindingen uit eerder onderzoek (bijv. Butler, 2010; McDaniel et al., 2012, 2013; Roediger & Butler, 2011), dat de gemiddelde prestatie *tijdens* het oefenen verbeterde naarmate er meer geoefend werd. Het lijkt er dus op dat de studenten wel enigszins profiteerden van herhaald oefenen, maar dat zij niet in staat waren om hun antwoorden goed te onderbouwen. Helaas kon het effect op transfer niet worden vastgesteld omdat de prestatie van studenten op de transfertaken extreem laag was (Engels: *floor effect*). Mogelijk heeft de feedback het effect van herhaald oefenen op leren tenietgedaan. Volgens recent onderzoek is feedback alleen nuttig wanneer studenten niet in staat zijn om tot het goede antwoord te komen en heeft het nauwelijks invloed wanneer zij hier wel toe in staat zijn (Kliegl et al., 2019; Pastötter & Bäuml, 2016; Storm et al., 2014). Het is dus mogelijk dat de groep studenten die maar één keer oefende – en het minst goed presteerde tijdens het oefenen – de feedback beter heeft verwerkt en daardoor even goed presteerde op de posttest als de andere groepen. Dit idee wordt ondersteund door de bevinding dat de studenten meer tijd besteedden aan de feedback na onjuiste antwoorden op de oefentaken dan na juiste antwoorden.

De studie in hoofdstuk 6 richtte zich op het identificeren van belemmeringen voor succesvolle transfer van vaardigheden om kritisch te denken. Wanneer transfer niet succesvol is, dan zou dit kunnen komen doordat studenten niet herkennen dat de aangeleerde kennis relevant is voor het nieuwe probleem, doordat ze de aangeleerde kennis niet uit hun geheugen kunnen ophalen of doordat ze de aangeleerde kennis niet kunnen toepassen op het nieuwe probleem (het driestappenmodel van transfer; Barnett

& Ceci, 2012). Er werden twee experimenten uitgevoerd om vast te stellen waar het transferprobleem uit eerdere onderzoeken van dit proefschrift door veroorzaakt zou kunnen zijn: één experiment met universitaire studenten in het laboratorium en één met hbo-studenten in de context van een vak. De studenten ontvingen eerst instructies over kritisch denken. Daarna oefenden zij met syllogismen – dit zijn redeneertaken waarbij je moet bepalen of een getrokken conclusie geldig is – in de context van domeinrelevante problemen. Deze keer werd de prestatie van de studenten gemeten op syllogismen (hiermee werd het *leren* gemeten), syllogismen in nieuwsberichten of artikelen (hiermee werd *nabije transfer* gemeten) en nieuwe taken met dezelfde onderliggende principes als syllogismen (hiermee werd *verre transfer* gemeten). De studenten werden ingedeeld in vier verschillende groepen en afhankelijk van de groep, ontvingen zij tijdens het maken van de transfertaken op de posttest (1) geen ondersteuning, (2) hints dat de principes uit de instructie relevant waren voor deze taken (ondersteuning in herkenning), (3) hints dat de principes uit de instructie relevant waren voor deze taken en de vraag om de opgedane kennis over die principes op te halen uit het geheugen (kennis ophalen) of (4) hints dat de principes uit de instructie relevant waren voor deze taken en een kort overzicht op papier van deze principes (ondersteuning in ophalen). Kortom, er werd in de condities verschillende ondersteuning voor de verschillende stappen in het transferproces aangeboden. Door de effecten van de condities te vergelijken, valt dan af te leiden waar zich problemen in het bereiken van transfer voordoen bij de studenten (vlg. Butler et al., 2013, 2017). Binnen de 'kennis ophalen' groep werd daarnaast gekeken of het vermogen van de studenten om kennis op te halen uit het geheugen gerelateerd was aan hun prestaties op de transfertaken. In beide experimenten verbeterden de prestaties van studenten op de leer- en nabije transfertaken na de instructie en het oefenen. Er werd zelfs een verbetering op de verre-transfertaken gevonden, maar de prestatie op deze taken was over het algemeen vrij laag dus er was nog veel ruimte voor verbetering. Een interessante bevinding was dat er geen verschillen waren tussen de vier groepen. De studenten waren dus niet geholpen bij de ondersteuning voor herkenning en ophalen. Dit suggereert dat de studenten in staat waren te herkennen dat de aangeleerde kennis relevant was voor de nieuwe taken en dat zij deze kennis konden ophalen uit het geheugen, maar moeite hadden met het toepassen van deze kennis op de nieuwe taken. Echter, de bevindingen uit de 'kennis ophalen' groep ondersteunen het idee dat de studenten alleen problemen hadden met het toepassen van de aangeleerde kennis niet volledig. De meeste studenten haalden namelijk niet alle relevante informatie op uit het geheugen. Bovendien werd er (via verkennende analyses) een matig tot hoog positief verband gevonden tussen de opgehaalde kennis en de prestaties op de transfertaken. Dit wijst erop dat problemen in het ophalen van kennis in elk geval ten dele ook een rol spelen bij niet-succesvolle transfer. Beschrijvende statistieken ondersteunen dit idee: de studenten die ondersteuning in ophalen kregen, presteerden beter (hoewel niet significant beter) dan de studenten in de andere groepen op de verre-transfertaken van de posttest in het eerste experiment en op de nabije-transfertaken in het tweede experiment.

Conclusie

Dit proefschrift werpt meer licht op de complexiteit van het bevorderen van het leren en de transfer van kritisch-denken-vaardigheden van studenten in het hoger onderwijs. De bevindingen in dit proefschrift benadrukken het belang van expliciete instructie over kritisch denken in combinatie met oefening op domeinrelevante problemen voor het leren van kritisch-denken-vaardigheden. Het lijkt dus waardevol om kritisch denken in te bedden in hoger onderwijs curricula en het expliciet aan bod te laten komen in (kritisch denken) vakken. Tevens werd uit dit proefschrift duidelijk dat generatieve verwerkingsstrategieën geen wondermiddel zijn om leren te verbeteren en transfer te bewerkstelligen. Hoewel ze voor sommige vaardigheden goed werken, lijken ze namelijk niet bij te dragen aan het verder bevorderen van de vaardigheid om systematische redeneerfouten te vermijden (bovenop de effecten van instructie en oefening). Daarnaast maken de studies in dit proefschrift eens te meer duidelijk hoe moeilijk het is om kritisch-denken-vaardigheden zodanig te trainen dat er ook *transfer* optreedt naar nieuwe situaties. Het lijkt zinvol om in (onderzoek naar) onderwijs in kritisch denken meer aandacht te besteden aan factoren die succesvolle transfer van kritisch-denken-vaardigheden kunnen belemmeren, zoals problemen met het ophalen van aangeleerde kennis uit het geheugen en met het toepassen van deze kennis in een nieuwe context. Met het oog op dat laatste, zou het interessant zijn om verder te onderzoeken hoe studenten ondersteund kunnen worden in de toepassing van kritisch-denken-vaardigheden.

Om studenten te helpen goede kritische denkers te worden – in de zin dat zij aangeleerde vaardigheden kunnen toepassen op verschillende taken en in verschillende contexten – lijkt het waardevol om in toekomstig (praktijkgericht) onderzoek de effecten van langere interventies of uitgebreidere cursussen gericht op kritisch denken te onderzoeken. Daarnaast is het van belang om in toekomstig onderzoek ook aandacht te besteden aan de ontwikkeling en de verbetering van de denkhouding van studenten. Want interventies die de kritisch-denken-vaardigheden van studenten verbeteren, zullen in de praktijk weinig zoden aan de dijk zetten wanneer studenten niet de juiste denkhouding hebben en geen mentale inspanning willen leveren om deze vaardigheden te gebruiken. Kortom, de vraag is hoe we studenten uit kunnen dagen om in kritisch denken te investeren.

Curriculum vitae

Curriculum vitae

Lara van Peppen was born in Delft, the Netherlands, on April 28, 1992. After completing her secondary education at the Stanislascollege in Delft in 2010, she started studying Psychology at Leiden University from which she obtained her bachelor's degree in 2013. Subsequently, she enrolled in the master's specialization Applied Cognitive Psychology at Leiden University from which she obtained her degree in 2015. For her master's thesis, which focused on the effects of physical load on cognitive performance, she did a research internship at the Training & Performance Innovations expertise group of the Netherlands Organisation for Applied Scientific Research (TNO). Following her interest in (applied) research, Lara became a PhD candidate at the Department of Psychology, Education and Child Studies at Erasmus University Rotterdam in January 2016, studying how to foster students' critical thinking skills in such a way that these would also transfer across tasks and contexts. Her project was part of the broader NWO-funded research project "Investing in Thinking Pays Good Interest: Improving Critical Thinking Skills of Students and Teachers in Higher Professional Education". During her PhD trajectory, Lara participated in the Brain and Learning research group of Avans University of Applied Sciences and collaborated with educational advisors, teachers, and researchers from Avans University of Applied Sciences and Utrecht University. Further, Lara was a visiting scholar in prof. dr. Patricia Alexander's Disciplined Reading and Learning Research Lab at the University of Maryland in College Park, MD, United States (September – November 2019). Whilst conducting her PhD research, Lara was a member of the Interuniversity Centre for Educational Sciences (ICO), presented her research at various (inter)national conferences and symposia, was the recipient of two best poster awards, worked as an academic teacher/trainer, co-supervised master's thesis projects, and organized several events, symposia, and meetings. Further, she gave workshops on various educational and psychological topics to students and educational professionals. Lara is currently working as educational innovator at the Erasmus University Medical Center in Rotterdam, focusing on blended learning and assessment and feedback in medical education.

Publications

Janssen, E. M., Mainhard, T., Buisman, R. S. M., Verkoeijen, P. P. J. L. Heijltjes, A. E. G., **Van Peppen, L. M.**, & Van Gog, T. (2019). Training higher education teachers' critical thinking and attitudes towards teaching it. *Contemporary Educational Psychology, 58*, 310–322. https://doi.org/10.1016/j.cedpsych.2019.03.007

Janssen, E. M., Meulendijks, W., Mainhard, T., Verkoeijen, P. P. J. L., Heijltjes, A. E., **Van Peppen, L. M.**, & Van Gog, T. (2019). Identifying characteristics associated with higher education teachers' Cognitive Reflection Test performance and their attitudes towards teaching critical thinking. *Teaching and Teacher Education, 84,* 139–149. https://dx.doi.org/10.1016/j.tate.2019.05.008

Janssen, E. M., Verkoeijen, P. P. J. L., Heijltjes, A. E. G., Mainhard, T., **Van Peppen, L. M.**, & Van Gog, T. (2020). Psychometric properties of the Actively Open-minded Thinking scale. Thinking Skills and Creativity, 36. https://doi.org/10.1016/j.tsc.2020.100659

Van Peppen, L. M., Verkoeijen, P. P. J. L., Heijltjes, A. E. G., Janssen, E. M., Koopmans, D, & Van Gog, T. (2018). Effects of self-explaining on learning and transfer of critical thinking skills. *Frontiers in Educational Psychology, 3*, 100. https://doi.org/10.3389/feduc.2018.00100

Vereijken, M. W. C., Baas, M., Van Dijk, E. E., Megawanti, M., **Van Peppen, L. M.**, Poort., I. Soppe, K. F. B., Tacoma, S., Wijbenga, M., & Van der Rijst, R. (2019, Oktober 7). *Reflections on contemporary trends in Dutch higher education research.* ICO Education. https://ico-education.nl/wp-content/uploads/2019/10/T9-Current-trends-in-research-into-higher-education-in-the-Netherlands-2.pdf

Papers

Janssen, E. M., Mainhard, T., Verkoeijen, P. P. J. L., Heijltjes, A. E. G., **Van Peppen, L. M.**, & Van Gog, T. (submitted). *Training higher education teachers' ability to explain biases in students' reasoning.* Manuscript submitted for publication.

Van Peppen, L. M., Van Gog, T., Verkoeijen, P. P. J. L., & Alexander, P. A. L. (submitted). *Identifying obstacles to transfer of critical thinking skills.* Manuscript submitted for publication.

Van Peppen, L. M., Verkoeijen, P. P. J. L., Heijltjes, A. E. G., Janssen, E. M., & Van Gog, T. (submitted). *Enhancing students' critical thinking skills: Is comparing correct and erroneous examples beneficial?* Manuscript submitted for publication.

Van Peppen, L. M., Verkoeijen, P. P. J. L., Heijltjes, A. E. G., Janssen, E. M., & Van Gog, T. (submitted). *Repeated retrieval practice to foster students' critical thinking skills.* Manuscript submitted for publication.

Van Peppen, L. M., Verkoeijen, P. P. J. L., Kolenbrander, S. V., Heijltjes, A. E. G., Janssen, E. M., & Van Gog, T. (submitted). *Learning to avoid biased reasoning: Effects of interleaved practice and worked examples.* Manuscript submitted for publication.

Verkoeijen, P. P. J. L., Koppenol-Gonzalez, G. V., **Van Peppen, L. M.**, Broeren, M. M. D. H. J., Heijltjes, A. E. G., Kuijpers, R. E., Nobelen, J. T. L. M., & Tillema, M., & Arends, L. R. *Assessing the generality of a self-administered strategic resource use intervention on academic performance: A multi-site, preregistered conceptual replication of Chen, Chavez, Ong & Gunderson.* Provisionally accepted for publication in *Advances in Methods and Practices in Psychological Science*.

Presentations

Scientific presentations on the studies that have been conducted as part of this research project to educational professionals and researchers at (inter)national conferences and symposia. Presenting author(s) indicated with *.

Van Peppen, L. M.*, Verkoeijen, P. P. L. J., Heijltjes, A. E. G., Janssen, E. M., & Van Gog, T. (2019, August). *Contrasting correct and erroneous examples to enhance students' critical thinking skills.* Oral presentation as part of the symposium 'Critical Thinking in Higher Education: Educational Guidelines and Instructional Interventions' (Organizers: L.M. van Peppen & E. M. Janssen) at the biannual conference of the European Association for Research on Learning and Instruction (EARLI 2019), Aachen, Germany.

Van Peppen, L. M.*, Verkoeijen, P. P. L. J., Heijltjes, A. E. G., Janssen, E. M., & Van Gog, T. (2019, March). *Enhancing students' critical thinking skills: Is contrasting correct and erroneous examples beneficial?* Pitch at the Graduate Research Day of the Department of Psychology Education and Child Studies (GRD DPECS) of the Erasmus University Rotterdam, Rotterdam, the Netherlands.

Van Peppen, L. M.*, Verkoeijen, P. P. L. J., Heijltjes, A. E. G., Janssen, E. M., & Van Gog, T. (2018, August). *Can contrasting correct and erroneous examples enhance students' critical thinking skills?* Poster presentation at the biannual conference of Special Interest Groups 6 and 7 (Instructional Design & Technology Enhanced Learning and Instruction) of the European Association for Research on Learning and Instruction (EARLI SIG 6-7 2018), Bonn, Germany.
Awarded with (1) Best Poster Presentation Award, EARLI SIG 6–7 conference 2018 and (2) Award for PhD Excellence: Best Poster 2018, Erasmus Graduate School of Sciences and the Humanities.

Van Peppen, L. M.*, Kolenbrander, S.V., Verkoeijen, P. P. L. J., Heijltjes, A. E. G., Janssen, E. M., & Van Gog, T. (2018, April). *Learning to avoid biased reasoning: Effects of interleaved practice and worked examples.* Oral presentation as part of the symposium 'Teaching critical thinking: Assessing and improving students' and teachers' reasoning skills' (Organizers: L.M. van Peppen & E. M. Janssen) at the annual conference of the American Educational Research Association (AERA 2018), New York, NY, USA.

Van Peppen, L. M.*, Verkoeijen, P. P. J. L., Heijltjes, A. E. G., Janssen, E. M., Koopmans, D., & Van Gog, T. (2017, September). *Effects of self-explaining on learning and transfer of critical thinking skills.* Oral presentation as part of the symposium 'Critical thinking in higher education: A closer look at teachers and students' (Organizers: L.M. van Peppen & E. M. Janssen) at the biannual conference of the European Association for Research on Learning and Instruction (EARLI 2017), Tampere, Finland.

Van Peppen, L. M.*, Verkoeijen, P. P. J. L., Heijltjes, A. E. G., Janssen, E. M., Koopmans, D., & Van Gog, T. (2017, August). *Fostering students' critical thinking skills: Effects of self-explaining on learning and transfer.* Oral presentation at the biannual pre-conference of the Junior Researchers of EARLI (JURE 2017), the European Association for Learning and Instruction, Tampere, Finland.

Van Peppen, L. M.*, Kolenbrander, S. V., Verkoeijen, P. P. L. J., Heijltjes, A. E. G., Janssen, E. M., & Van Gog, T. (2017, May). O*ptimizing critical thinking instruction: What practice schedules and task-types are effective?* Poster presentation at the annual convention of the Association of Psychological Science (APS 2017), Boston, MA, USA.
Nominated for the Award for PhD Excellence: Best Poster 2017, Erasmus Graduate School of Sciences and the Humanities.

Van Peppen, L. M.*, Kolenbrander, S. V., Verkoeijen, P. P. L. J., Heijltjes, A. E. G., Janssen, E. M., & Van Gog, T. (2017, May). *Optimizing critical thinking instruction: What practice schedules and task-types are effective?* Poster presentation at the Interuniversity Center for Educational Sciences National Spring School (ICO NSS 2017), Utrecht, the Netherlands.

Kolenbrander, S. V.*, **Van Peppen, L. M.***, & Janssen, E. M.* (2017, April). *Kritisch onderwijzen [Teaching critically]*. Invited evening lecture at the Academielezing of Avans University of Applied Sciences, Breda, the Netherlands.

Van Peppen, L. M.*, Verkoeijen, P. P. J. L., Heijltjes, A. E. G., Janssen, E. M., Koopmans, D., & Van Gog, T. (2016, November). *Effects of self-explaining on learning and transfer of critical thinking skills.* Poster presentation at the annual conference of the European Association for Practitioner Research on Improving Learning (EAPRIL 2016), Porto, Portugal.

Van Peppen, L. M.*, Verkoeijen, P. P. J. L., Heijltjes, A. E. G., Janssen, E. M., Koopmans, D., & Van Gog, T. (2016, Oktober). *Effects of self-explaining on learning and transfer of critical thinking skills.* Poster presentation at the Interuniversity Center for Educational Sciences International Fall School (ICO IFS 2016), Bad Schussenried, Germany.

Workshops and training sessions

Workshops and training sessions on various educational and psychological topics to students, teachers, educational practitioners, and researchers. Presenting author(s) indicated with *.

Van Peppen, L. M.*, & Janssen, E. M.*, (2019, November). "*Van mening veranderen is een teken van zwakte": Wat is een kritische denkhouding en (hoe) kun je die bevorderen? ["Changing your opinion is a sign of weakness": What is a critical thinking attitude and (how) can you improve it?]*. Workshop at the symposium 'Laat je (niets) wijsmaken!' over kritisch leren denken in het hbo [symposium on critical thinking in higher professional education], Avans Hogeschool, Den Bosch, the Netherlands.

Raaijmakers, L. H.*, **Van Peppen, L. M.***, Tillema, M.*, & Van Harsel, M. (2019, February). *Motiverend lesgeven [Teaching motivating]*. Workshop at the Academie voor Industrie en Informatica (AI&I) of Avans University of Applied Sciences, Den Bosch, the Netherlands.

Van Peppen, L. M.* (2016 – 2019). *Kritisch leren denken [Learning to think critically]*. Numerous training sessions for students of Avans University of Applied Sciences and the Erasmus University Rotterdam in the context of research.

Van Peppen, L. M.*, & Van Harsel, M.* (2018, November). *Effectief leren en studeren: hoe doe je dat? [Learning and studying effectively: How to do that?].* Workshop at Beroepshavo MBO College Hilversum, the Netherlands.

Janssen, E. M.*, & **Van Peppen, L. M.*** (2017, November). *Kritisch denken doceren kun je leren?! [Teaching critical thinking can be learned?!]*. Workshop during the expertmeeting 'Werk maken van kritisch denken in het hbo' of the Vereniging Hogescholen on critical thinking in higher professional education, Driebergen, the Netherlands.

Heijltjes, A. E. G.*, Janssen, E. M.*, & **Van Peppen, L. M.*** (2016, Oktober). *Kritisch denken loont de moeite [Critical thinking pays off]*. Workshop at the symposium 'Leren in het hbo: denken, doen en laten', organized by the Brain and Learning research group of Avans University of Applied Sciences, Breda, the Netherlands.

Dankwoord

Acknowledgements in Dutch

De zon schijnt de kamer in, op de achtergrond hoor ik Marcus Mumford zingen "I lift up my eyes to a new high". In gedachten neem ik de afgelopen jaren door en ik besluit de allerlaatste hand aan mijn proefschrift te leggen. Het is er de tijd voor. De afgelopen jaren vormden een prachtig leerzame periode. Dankbaar ben ik eenieder die mij geholpen, aangemoedigd of gesteund heeft. Dankzij jullie was het voor mij mogelijk dit proefschrift te schrijven en me op persoonlijk vlak te ontwikkelen. In het bijzonder wil ik de volgende personen bedanken.

Allereerst dr. Peter Verkoeijen, prof.dr. Tamara van Gog en dr. Anita Heijltjes voor de begeleiding tijdens mijn promotietraject. Het was een eer om jullie als mijn begeleidingsteam te hebben.

Peter, jij hebt me op zoveel vlakken dingen geleerd. Wekelijks maakte je tijd om onder het genot van een kop koffie bij te praten over de dagelijkse gang van zaken en om inhoudelijk te discussiëren. Van ontwikkelingen binnen het onderwijs en vernieuwingen binnen de wetenschap tot het spel van NAC Breda en het ongemak bij Boer zoekt Vrouw. Jouw uiterst deskundige en altijd snelle commentaar, stimuleerde me om het onderste uit de kan te halen. Je gaf me vertrouwen en verschafte me de inzichten die ik nodig had om op eigen benen te gaan staan. Ik heb jouw wijze raad en interesse in mijn leven naast het onderzoek altijd zeer gewaardeerd. Dat we in de toekomst nog maar geregeld een kop koffie met elkaar mogen gaan drinken.

Tamara, jij wist me met jouw scherpe blik uit te dagen mijn eigen werk kritisch te beschouwen. Jouw theoretisch denkvermogen en waardevolle inzichten hebben dit proefschrift naar een hoger niveau getild. Naast het begeleiden van mijn onderzoek was je ook altijd geïnteresseerd in mijn (carrière)ontwikkeling. Bedankt voor je oprechte betrokkenheid, het vertrouwen dat je in me had en de vele kansen die je me geboden hebt.

Anita, het is grotendeels aan jouw inspanningen te danken dat dit proefschrift hier ligt. Met jouw promotieonderzoek en opgebouwde kennis over het onderwijzen van kritisch denken, heb je feitelijk het fundament gelegd. Al mijn stukken werden door jou grondig gelezen en van inspirerende suggesties voorzien. Ik heb veel bewondering voor de wijze waarop jij wetenschap naar praktijk weet te vertalen. Bedankt voor de integere en persoonlijke manier van begeleiden, het was een voorrecht om van jou te mogen leren.

De leden van de promotiecommissie, prof.dr. Fred Paas, prof.dr. Paul van den Broek en dr. Katinka Dijkstra dank ik voor het kritisch lezen en beoordelen van dit proefschrift en het deelnemen aan de oppositie. Tevens dank ik de overige leden van promotiecommissie, prof.dr. Jan Elen, prof.dr. Sofie Loyens en dr. Marion Tillema voor hun bereidheid om met mij van gedachten te wisselen over de inhoud van mijn proefschrift.

Et al., het is de afkorting in het Latijn voor 'en anderen'. In de wetenschap duidt het op degenen die een bijdrage hebben geleverd aan een product. Alleen al door de afkorting wordt de bijdrage van deze personen vaak onderschat. Dit proefschrift is veel meer dan een product van mij alleen. Het is tot stand gekomen dankzij de samenwerking en de steun van vele collega's. Graag richt ik me hier dan ook tot deze 'anderen'.

Dankbaar ben ik dat ik onderdeel mocht zijn van het project 'Investing in Thinking Pays Good Interest'. Een project waarbij kritisch denken en leren in de praktijk centraal staan, dat vraagt natuurlijk om eenzelfde houding van de betrokkenen. Anita, Eva, Peter, Tamara en Tim, bedankt voor al jullie waardevolle ideeën en kritische kanttekeningen, maar bovenal voor de fijne sfeer tijdens onze bijeenkomsten. Ondanks de bovengemiddelde affiniteit met de Duitse taal en menig grap die daardoor aan mij voorbij is gegaan, viel er met jullie heel wat te lachen. In het bijzonder wil ik mijn medepromovenda Eva bedanken, ik denk met een lach op mijn gezicht terug aan onze samenwerking, gesprekken, borrels en congresbezoeken. Met onze verschillende karakters wisten wij elkaar goed aan te vullen en ik hoop dat onze wegen zich in de toekomst nog veel vaker zullen kruisen. Dat de lijnen Rotterdam-Utrecht-Breda maar kort mogen blijven!

Ook de vele collega's van de Erasmus Universiteit bij wie ik terecht kon voor advies en momenten van ontspanning ben ik erkentelijk. Met name de collega's uit de O&O-sectie en alle medepromovendi wil ik bedanken voor de collegiale sfeer en de gezellige lunchpauzes. Een aantal in het bijzonder, Donna, Eke, Gertjan, Işıl, Jacqueline, Jason, Keri, Loïs, Marloes, Milou, Miranda, Sabrina en Willemijn, niets is zo fijn als samen in hetzelfde schuitje zitten. Ilse, bedankt voor de fijne gezamenlijke start en onze aangename momenten samen. Denise, Iris, Lara en Marieke, jullie oprechte interesse maakte dat ik me al snel thuis voelde en droeg eraan bij dat ik met plezier naar werk kwam. Anniek en Julia, het was fijn om de dag te starten met een (langer dan gepland) praatje in jullie kamer. Rob, bedankt dat je altijd tijd maakte voor een kop koffie en een portie droge humor. Ik hoop dat we die momenten erin blijven houden. Onvergetelijk was ook mijn kamergenoot Marijntje, ik ben dankbaar dat ik vier jaar naast jou heb mogen doorbrengen.

Alle collega's van Avans Hogeschool en het lectoraat Brein & Leren wil ik bedanken voor de fijne donderdagen vol interessante discussies en gezelligheid. Anita, Anton, Eva, Hans, Ilse, Janneke, Lottie, Marion, Marloes, Michael, Milou, Peter, Stefan, Suzan en Yvonne, jullie passie voor het onderwijs is aanstekelijk en heeft ervoor gezorgd dat ik altijd oog voor de praktijk hield. In het bijzonder wil ik Stefan bedanken, ik was nog maar net gestart of jij had al een groep studenten paraat voor mijn onderzoek. Het was de start van een vruchtbare, maar bovenal ontzettend prettige samenwerking. Marion, ik heb bewondering voor de manier waarop jij wetenschap en praktijk weet samen te brengen en ben vereerd dat je in mijn promotiecommissie wilt plaatsnemen. Marloes, Milou en Suzan, bedankt voor de fijne gesprekken en wijze adviezen. Aan allen hierboven, houdoe en bedankt!

I would like to sincerely thank prof.dr. Patricia Alexander for the opportunity to visit her lab at the University of Maryland. Patricia, it was a true honor working with you and learning as much as I did from you. Also, I would like to thank the other members of the 'Alexander family' for all the great conversations and discussions. Special thanks to Anisha, Eric, Julianne, and Yuting, for making me feel so welcome during my stay. Eric and Julianne, I very much enjoyed our days working in the sun (while drinking cappuccino *vanilla*) and our dinners. To all, hope we'll meet again!

Een woord van dank aan alle studenten die hebben deelgenomen aan het onderzoek. Esther en Marjolein, veel dank voor jullie hulp als student-assistent. Ik hoop dat jullie geen blijvende weerzin tegen 'als-dan-beredeneringen' hebben.

Mijn paranimfen en goede vriendinnen, Iris en Simone. Wat ben ik blij dat jullie naast mij staan.

Iris, tijdens mijn zoektocht als beginnende promovenda kwam jij op mijn pad. Hoe toepasselijk dat jij, met jouw Griekse achtergrond, mijn promotietraject richting gaf, zoals de Griekse filosofen vormgaven aan het onderwerp van mijn proefschrift: kritisch denken. Bedankt voor al je wijze advies, het delen van onze vele vertwijfelingen (stelling 6 zal ook jou wel bekoren, denk ik), alle plezierige tijd die we samen hebben doorgebracht en je vriendschap. Zonder jou was het nooit zo leuk geweest.

Simone, een toevallige ontmoeting tijdens onze allereerste les in statistiek precies 10 jaar geleden en nu staan we hier. In al die jaren is onze vriendschap alleen maar sterker geworden. Bedankt voor jouw puurheid, het vertrouwen dat je me geeft en de vele dierbare herinneringen. Laten we samen nog jaren om dezelfde grappen blijven lachen. 'Baie dankie' dat je er altijd voor me bent.

Op persoonlijk vlak is er een aantal mensen dat me tijdens het schrijven van dit proefschrift heeft bijgestaan. Zij gaven mij het kostbare gevoel dat mijn leven uit veel meer bestond dan werk.

Denise, Janneke, Nadine, Samira en Thari, bedankt voor jullie begrip als ik even wat minder tijd had en de onvoorwaardelijke vriendschap. Debby en Diede, onze altijd gezellige etentjes, picknicks en andere uitstapjes gaven me de broodnodige ontspanning. Annelies, Eline en Mariska, de een nog langer in mijn leven dan de ander, ik ben blij dat wij elkaar altijd zullen treffen. Teamgenoten, bedankt voor de ontspanning tijdens de wedstrijden en de gezelligheid na afloop.

Een speciaal woord van dank aan mijn gehele (schoon)familie voor hun rotsvaste steun en toeverlaat. Bedankt voor jullie interesse in mijn proefschrift en de vele mooie momenten samen. Opa en oma, wat bijzonder dat ik jullie mijn proefschrift kan overhandigen. Tim, met jouw ondernemende en creatieve geest weet je me vaak te inspireren en ben je op vele facetten van mijn promotietraject van invloed geweest. Dankbaar ben ik dat jij, zoals het een oudere broer betaamt, altijd voor mij klaar staat.

Mijn ouders, Roland en Marianne, ik kan jullie niet genoeg bedanken voor jullie onvoorwaardelijke steun en liefde. Van kinds af aan hebben jullie me gestimuleerd om het beste uit mezelf te halen en mijn hart te volgen. Jullie hebben de basis gelegd en dat ik dit met jullie kan delen, geeft glans aan het geheel. Bedankt dat jullie er altijd voor mij zijn.

Tenslotte Nick, bedankt voor alle steun en vrijheid die je me gegeven hebt. Met jouw positieve levensinstelling, relativeringsvermogen, gevoel voor humor en liefde, lever jij een onmisbare bijdrage aan mijn leven. Ik ben blij dat jij er bent.

Lara, juli 2020

Alles blijft
Alles gaat voorbij
Alles blijft voorbijgaan
— Jules Deelder