**At the heart of learning: navigating towards educational neuroscience in health professions education**
Versteeg, M.

**Citation**
Versteeg, M. (2021, January 21). *At the heart of learning: navigating towards educational neuroscience in health professions education*. Retrieved from https://hdl.handle.net/1887/3134566

| | |
|---|---|
| Version: | Publisher's Version |
| License: | Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden |
| Downloaded from: | https://hdl.handle.net/1887/3134566 |

**Note:** To cite this publication please use the final published version (if applicable).

Cover Page

Universiteit Leiden

Leiden University Repository

The handle https://hdl.handle.net/1887/3134566 holds various files of this Leiden University dissertation.

**Author**: Versteeg, M.
**Title**: At the heart of learning: navigating towards educational neuroscience in health professions education
**Issue Date**: 2021-01-21

# At The Heart Of Learning

Navigating towards educational neuroscience in health professions education

Marjolein Versteeg

**Colophon**

Cover design: Lars Rietkerk
Layout: Lars Rietkerk & Marjolein Versteeg
Printed by: Ridderprint
ISBN: 978-94-6416-234-9

This thesis was printed with financial support from Leiden University Medical Center and the Dutch Association for Medical Education (NVMO).

© Marjolein Versteeg, 2020
All rights reserved. No part of this thesis may be reproduced or transmitted in any form or by any means without permission of the copyright owner.

# At The Heart Of Learning

Navigating towards educational neuroscience in health professions education

Proefschrift

Ter verkrijging van

de graad van Doctor aan de Universiteit Leiden,

op gezag van Rector Magnificus prof.mr. C.J.J.M. Stolker,

volgens besluit van het College voor Promoties

te verdedigen op donderdag 21 januari 2021

klokke 15.00 uur

door

Marjolein Versteeg

geboren te Hengelo

in 1994

**Promotor**

Prof. dr. P. Steendijk

**Copromotor**

Dr. A.J. de Beaufort

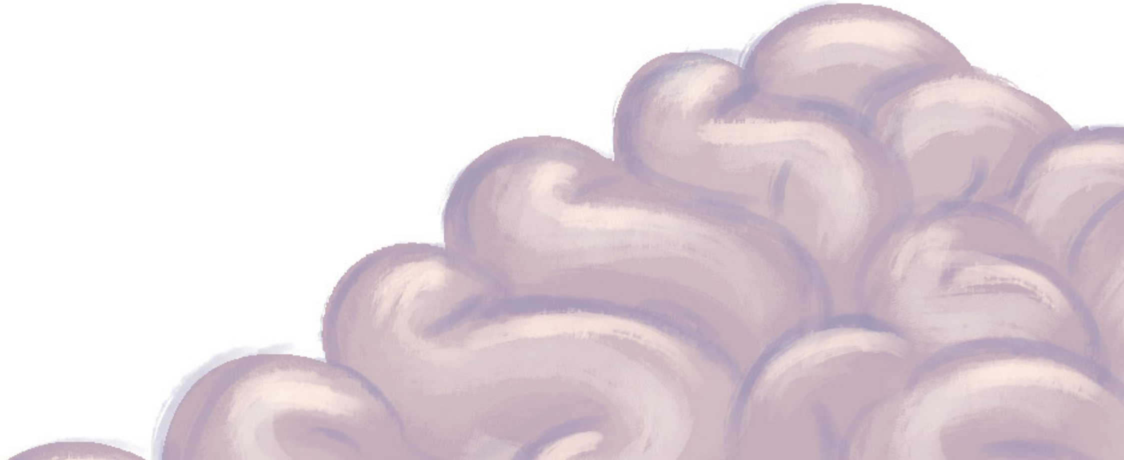
**Promotiecommissie**

Prof. dr. M.J. Schalij, Leiden University Medical Center
Prof. dr. F.W. Dekker, Leiden University Medical Center
Prof. dr. M. Wijnen-Meijer, Technical University of Munich
Prof. dr. A.B.H. de Bruin, Maastricht University

## Table of contents

# Chapter 1

# General introduction



## · Scope ·

Learning influences the brain. Every time you learn a new fact, a new concept or a new skill your brain has changed. The brain also influences learning. Your ability to learn is constrained by the architecture and functioning of the brain. Despite the close relationship between the brain and learning, neuroscience has remained remarkably distant from the classroom. In my quest to improve health professions education, this gap highlights an opportunity. Building a bridge between neuroscience and education may pave the way for evidence-informed education in the health professions.

In this thesis, I use *educational neuroscience* as a source of inspiration for developing evidence-informed health professions education. The overall aim of this thesis is to improve health professions education by investigating learning processes using an educational neuroscience-inspired approach.

This introductory chapter describes the rationale behind our studies. I describe the scientific field of educational neuroscience as well as its relevance to health professions education. Subsequently, I elaborate on three specific learning processes; spaced learning, concept learning, and metacognitive learning, as these are the focus of our research. This chapter concludes with an overview of the studies included in this thesis.

## · The science of learning and why it matters ·

*"When the why is clear, the how is easy."* – J. Rohn

For decades, there has been a drive towards evidence-informed health professions education. Evidence-informed implies that education should be integrated with evidence from research to find best practices and improve the quality of education and healthcare accordingly (Nelson & Campbell, 2017; Sharples, 2013). Educators and policymakers are increasingly called upon to apply evidence-informed education in their curricula in order to facilitate meaningful and effective education (Thomas et al., 2019a; Durning et al., 2012; Van Der Vleuten, 2000).

The pursuit of educational excellence should be reflected by curricula with a scientifically sound basis (Ramani, 2006). Research informing our curricula should partly be directed at advancing our understanding of students' learning processes (Ruiter et al., 2012). For example, what they learn and how they learn. Research on learning processes can be informed by the tremendous growth in knowledge about the human brain over the last thirty years (Ansari et al., 2011). And so, I arrive at the interdisciplinary research field named educational neuroscience, which connects neuroscience, through cognitive psychology, with education sciences.

Why educational neuroscience? To become more evidence-informed, health professions education should start looking beyond boundaries. The evidence on the science of learning in fields such as neurobiology, cognitive psychology and higher education is there for the taking. My goal is to blur the boundaries between these fields using educational neuroscience-inspired research, in order to achieve: more evidence-informed education, more effective education, better learners, better healthcare professionals, and ultimately, better healthcare.

## · Educational neuroscience ·

*"Integrated research of two domains should be driven by an urge to understand." – A. Einstein*

There is a growing belief among scientists, teachers, and policymakers that education benefits from an understanding of the brain (Jones, 2009). Early enthusiasts included educational psychologists such as Edward Lee Thorndike. In his PhD thesis written in 1926, dr. Thorndike already claimed that learning has its physiological basis in the structure and activities of neurons and accessory organs which compose the nervous system (Thorndike, 1926). The field of educational neuroscience established itself several decades later, in the early 90s.

Educational neuroscience is an interdisciplinary scientific field which combines research from cognitive neuroscience, cognitive psychology, and other related disciplines to explore the interactive processes between biology and education (Goswami, 2006). The collective belief is that these disciplines should act in synergy in order to understand the learning brain and improve education (Sigman et al., 2014). As written in a recent commentary by Howard-Jones et al. (2016):

> *"The relationship between neuroscience and educational practice can be likened to the relationship between molecular biology and drug discovery, including the arduous process of clinical trials. The basic science tells you where to look, but does not prescribe what to do when you get there. Similarly, neuroscience may tell you where to look – that is, what neural functions are typical or impaired and how these operate – but this knowledge must be transformed by pedagogical principles and then assessed by behavioural trials in educational contexts, the equivalent of clinical drug trials."*

It is important to emphasise that our research does not aim to improve education through neuroscience directly. As illustrated by the quote, educational neuroscience utilises neuroscience as a supplementary ground which can anchor and enrich educational practice. If a psychological construct has a biological substrate, it will be better understood if the underlying mechanisms are supported by both behavioural and biological data. In turn, a better understanding of behavioural and biological processes leads to better guidance for educational interventions (Ansari et al., 2011; Howard-Jones et al., 2016). Vice versa, educational practice may serve as a source of inspiration for neuroscientists, by providing novel research conditions through unique real-world settings (Goswami, 2006; Sigman et al., 2014). Our research is inspired by the systematic interactions between neuroscience, psychology, and education on the fundamental and practical level. These interactions may lead to a common language with common questions to advance theory and practice in health professions education (Thomas et al., 2019b).

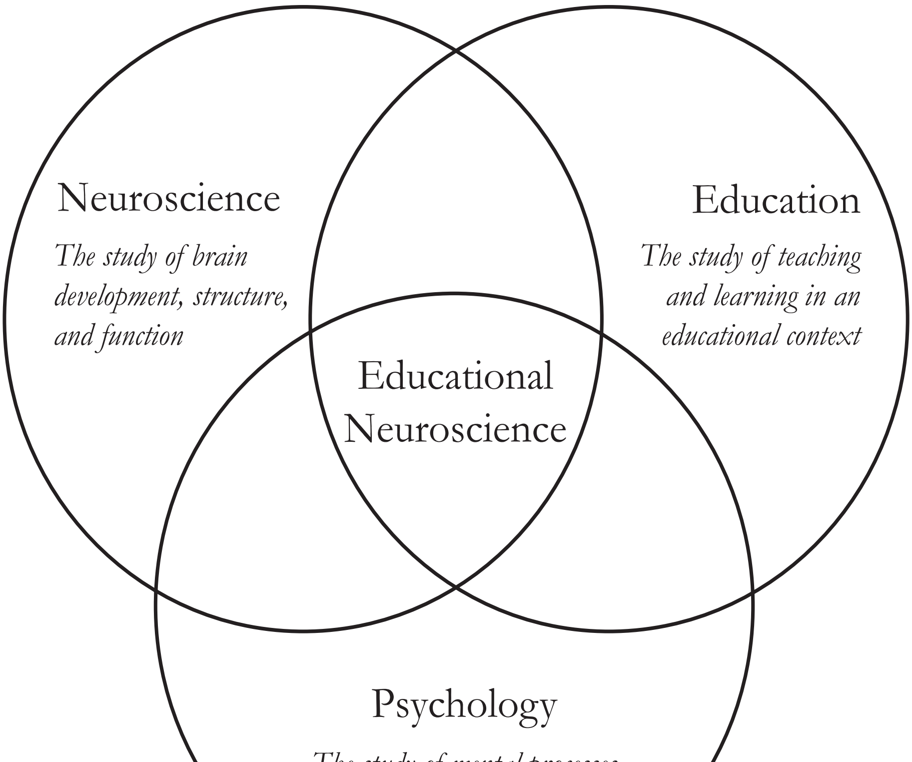



Figure 1 | The systemic interactions between neuroscience, education and psychology. Adapted from Sousa, 2010.

## · Focusing the research topic ·

*"The sun's rays do not burn until brought to a focus." – A. Bell*

The current role of educational neuroscience in health professions education research is limited (de Bruin, 2016; Friedlander et al., 2011; Ruiter et al., 2012). Nevertheless, various scientists have claimed that neuroscientific theories may have great impact on health professions education (Patel et al., 2009; Regehr & Norman, 1996). This includes translating insights from neuroscience to education on the one hand, and informing the science of learning through educational best-practices on the other hand. This line of reasoning has led to the creation of this thesis.

In this thesis, we focus on studying learning processes. But, how do we decide which specific learning processes should be studied? Considerable opportunities for health professions education have already been discussed in the literature. We selected three highly interesting research topics from the viewpoint of health professions education, which are also actively studied in neuroscientific and psychological disciplines: 1) spaced learning, 2) concept learning, and 3) metacognitive learning (de Jong et al., 2009; Friedlander et al., 2011; Ruiter et al., 2012). Together, these learning processes are fundamental to the three types of knowledge described by Bloom's revised taxonomy for establishing educational goals (Herwaarden et al., 2009; Krathwohl, 2002). First, spaced learning is a process by which learning is distributed over time in order to enhance retention. From Bloom's perspective, spaced learning facilitates acquirement of factual knowledge, which is necessary for health professionals to establish a solid knowledge base. Second, concept learning is a process related to the understanding of scientific concepts. From Bloom's perspective, concept learning facilitates conceptual knowledge, which is essential to health professionals in understanding the mechanisms of human body function. Third, metacognitive learning is a process that focuses on reflexivity; "thinking about one's thinking". From Bloom's perspective, metacognitive learning facilitates metacognitive knowledge, which is of critical importance for achieving the desired lifelong learning attitude in our health professionals. Below, I provide an outline of each learning process and its relevance to health professions education.

## · Spaced learning ·

*"There is no learning without remembering." - Socrates*

Students have a hard time recalling the learning material taught in medical school. Recent studies have demonstrated that half of the first-year medical knowledge could not be reproduced by second-year medical students who were retested unprepared (Schneid et al., 2019; Weggemans et al., 2017). Medical educators and students should be aware of this, since long-term retention of knowledge is of great importance for accurate clinical reasoning and adequate clinical practice. Consequently, there is a need for methods that improve knowledge retention.

Spaced learning is a method that could help to amend the difficulty of knowledge retention. If learning is distributed over multiple sessions and repeated over time this leads to better and longer retention (Carpenter et al., 2012). In health professions education, innovative forms of spaced learning are finding their way to the classroom. However, an overview of these applications is lacking due to the great diversity in the terms and definitions used in the literature to refer to spaced learning. Moreover, in the light of educational neuroscience we are interested if such research is informed by psychological or neuroscientific theories. For example, neuroscientists have recently advocated the use of short timescales in spaced learning (Smolen et al., 2016). Their research shows that biochemical cascades involved in memory formation act on different temporal domains with timescales from seconds to hours to days. The dynamics of molecules working on these timescales, such as second messengers and kinases, may contribute to the spacing effect. It would be of interest to investigate if short spaces could be of use in those settings that are currently overlooked, such as traditional lectures.

Spaced learning is the subject of study in Chapter 2 and Chapter 3 of this thesis. These chapters provide an overview of spaced learning studies in health professions education research and the use of psychological and neuroscientific theories. Additionally, the potential of using short spaces in a spaced lecture format is explored.

## · Concept learning ·

*“The more you know, the more you can know.” - Aristotle*

Students find it challenging to obtain an accurate conceptual understanding of human physiology (Michael, 2007). Their level of conceptual understanding has shown to be rather limited and difficult to enhance through traditional teaching methods (Palizvan et al., 2013). One of the primary reasons may be that students suffer from misconceptions, which impact on their learning process. Misconceptions can be defined as ideas that are incorrect according to current scientific views, resulting in misunderstanding of new information (Wandersee et al., 1994). Misconceptions appear very resistant to change, since they continue to exist even after taking the corresponding courses at university (Palizvan et al., 2013). Research that investigates best-practices on how to alleviate misconceptions in health professions education is scarce.

Different educational and neuroscientific theories exist on the process of alleviating misconceptions, which is referred to as conceptual change. Conceptual Change Theory describes conceptual change as shifting away from a misconception towards the scientifically correct conception (Posner et al., 1982). Conceptual change is a process of accommodation during which misconceptions are reorganised or replaced by the scientific conception. However, no study found that a particular learner's conception was completely extinguished and replaced by the current scientific view (Duit & Treagust, 2012). Accordingly, neuroscientists have suggested that old ideas stay alive as they can be used in particular contexts (Mareschal et al., 2013). In the light of educational neuroscience, it would be interesting to investigate how concept learning comes about to be able to inform health professions education.

Concept learning is the subject of study in Chapter 4 to Chapter 6 of this thesis. In sum, these chapters describe explorative research on the origins of misconceptions, and an educational intervention aiming to enhance concept learning among students. The study described in Chapter 6 moves from educational practice to the neuroscientific laboratory. Here, educational research marries cognitive neuroscience to challenge the above-mentioned theories on concept learning.

## · Metacognitive learning ·

*"I am thankful for the brain that was put in my head. Occasionally, I love to just stand to one side and watch how it works." – R. Bolles*

Metacognition is referred to as thinking about one's thinking (Flavell, 1979). Metacognitive learning as a component of self-regulated learning is gaining attention in our research community (Brydges & Butler, 2012; Gooding et al., 2017). Students are expected to become self-regulated learners which allows them to learn independently, effectively and lifelong (Group, 1996; Murdoch-Eaton & Whittle, 2012). Explicit teaching of metacognitive skills may help students to become more self-regulatory (Bjork et al., 2013). These skills include planning, monitoring, and evaluating one's actions (Zohar & Barzilai, 2013). However, explicit teaching of metacognitive skills in health professions education is scarce (Artino Jr et al., 2012). An additional problem is the low level of students' metacognitive knowledge, meaning they often do not know what they (do not) know (Thiede et al., 2003). This can become problematic for learning physiology concepts in particular, where misconceptions are often present unconsciously. Despite the recognised importance of metacognitive learning, research on metacognition with a focus on enhancing conceptual understanding is rather limited in health professions education.

Since metacognitive processes are suggested to be partly unconscious, they are difficult to investigate. Evidence is gathered by neuroscientists on brain areas that seem to drive metacognition, but there is still a large unknown area to be explored (Chua et al., 2014; Fleming et al., 2014). In the light of educational neuroscience, it would be of interest to gain insight in students' metacognition to guide future neuroscientific research.

Metacognitive learning is the subject of study in Chapter 7 to Chapter 10 of this thesis. The majority of studies focus on students' level of metacognitive knowledge. The final study, described in Chapter 10, maps students' metacognitive skills and perceptions on self-regulated learning in their medical curriculum.

## · Through the lens of critical realism ·

*"When you change the way you look at things, the things you look at change." – W. Dyer*

To understand the nature of the research in this thesis one must realise that a scientist, including myself, always uses a specific lens while conducting research. This lens is one's research paradigm.

The bridge between neuroscientific research and educational research is difficult to cross, due to differences in philosophies about learning (Flobakk, 2015). One can look at learning as individual biological changes at a cellular level of the brain or rather as a social activity taking into account social interactions and the importance of context. The biological perspective is in line with the positivist paradigm of research, suggesting there is one true reality that can be observed. Positivism gives less consideration to social influences compared to existential aspects such as biochemical processes in the brain. On the other hand, the social perspective is in line with the constructivist paradigm of research, suggesting there are multiple realities that are constructed by people. Constructivism has a primary focus on social interactions and contextual features.

As I aim to emphasise both biological and social concepts, I take a post-positivism approach towards research (Bergman et al., 2012). The post-positivism paradigm suggests there is one truth, but it can never be truly observed (Phillips et al., 2000). This paradigm includes the ontological principles of critical realism (Collier, 1994). Critical realism provides a philosophical perspective that emphasises both social concepts and biological concepts, but without reducing one to the other. This is important, as educational neuroscience has an interdisciplinary endeavour at its heart (Flobakk, 2015). Educational neuroscience addresses questions that lay on the border between the social and the biological, so both biological and social explanations should be considered relevant. The philosophical position of critical realism can be located in-between positivism, viz. neurobiological mechanisms are predominant. and constructivism, viz. social context is predominant (Collier, 1994; Flobakk, 2015).

Regarding educational neuroscience, critical realism allows for different methods of studying learning that align with different perspectives of the biological and social world. In this thesis, I seek for ways to understand and improve learning, by considering the influences of biological processes and the influences of individual perceptions and contexts.

## · Setting the scene ·

*"I welcome truth. But I wish all of my facts to be in their proper context." – G. Hinckley*

All studies in this thesis were conducted in the undergraduate curriculum at the Leiden University Medical Center. The 3-year program is a traditional curriculum mainly consisting of lectures, working groups and practical sessions. The study population comprises undergraduate students from different health professions: medicine, biomedical sciences, and clinical technology. In medicine, students focus on health prevention, and diagnosis and treatment of disease. In biomedical sciences, students specialise in cellular and molecular mechanisms of health and disease. In clinical technology, students learn about the technical aspects of healthcare. The undergraduate curricula have a strong focus on knowledge construction. Some studies in this thesis were carried out during undergraduate courses; 'mechanisms of disease' (Chapter 3), 'basis to homeostasis' (Chapter 4, 5, 9), 'human biology' (Chapter 7), and 'physiology, basic concepts' (Chapter 8). All studies took place between November 2016 and September 2019.

## · Aim and outline of this thesis ·

The overall aim of this thesis is to improve health professions education by investigating spaced learning, concept learning and metacognitive learning using an educational neuroscience-inspired approach.

For spaced learning, we investigated how spaced learning is currently implemented in health professions education. Particularly, what spacing formats are being used? And do short spaces benefit knowledge retention?

For concept learning, we investigated how conceptual change comes about. Particularly, what are the origins of students' misconceptions? Are there effective instructional designs that may enhance conceptual understanding in medical physiology education? And is conceptual understanding mainly a matter of conceptual change or conceptual shift?

For metacognitive learning, we investigated students' use of metacognition, specifically in the context of physiology education where conceptual understanding plays a prominent role. How can we assess students' metacognitive evaluation?

Do students use metacognitive skills, i.e. planning, monitoring, evaluating, while solving physiology questions? And how do students perceive self-regulated learning in their medical school curriculum?

To fulfil our aims, we conducted nine studies with specific research questions, of which an outline is provided in Table 1. Chapter 11 summarises the main findings of the work outlined in this thesis. Additionally, findings are discussed and implications for educational practice and future research are provided.

Table 1 | Studied research questions and corresponding research methodology.

| | Chapter | Research question | Research method |
|---|---|---|---|
| Spaced learning | 2 | How is spaced learning defined and applied in health professions education? | Scoping review |
| | 3 | Does the implementation of short spaces in a lecture enhance knowledge retention in students? | Experiment |
| Concept learning | 4 | What are the specific origins of inaccurate conceptual understanding among students regarding the interrelated concepts of pressure, flow, and resistance? | Content analysis |
| | 5 | Is cognitive inhibition involved in overcoming a physiological misconception? | Experiment |
| | 6 | Can peer instruction enhance students' comprehension of physiological concepts? | Experiment |
| Metacognitive learning | 7 | Can post-decision wagering be used as a measure of self-perceived knowledge in an educational context? | Experiment |
| | 8 | Can a multitier approach determine students' level of conceptual understanding by assessing their metacognitive evaluation? | Experiment |
| | 9 | Can refutation texts enhance students' cognition and metacognition regarding physiological concepts? | Experiment |
| | 10 | What are students' metacognitive competencies and what are their perceptions of self-regulated learning in the medical curriculum? | Thinking aloud & semi-structured interviews |

# Part 1

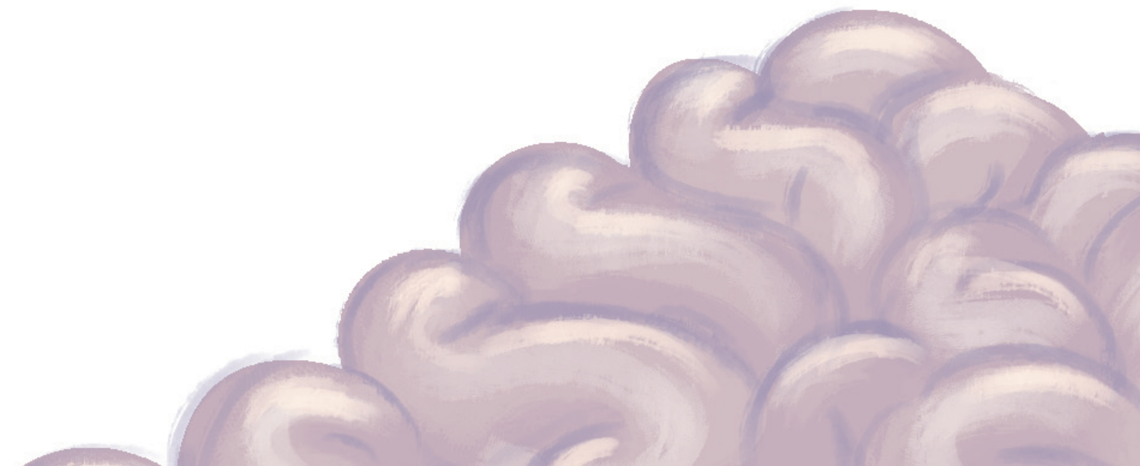

Spaced
Learning

# Chapter 2

# Conceptualising spaced learning in health professions education: a scoping review

Marjolein Versteeg · Renée A. Hendriks · Aliki Thomas ·
Belinda W. C. Ommering · Paul Steendijk

Published in: Medical Education (2019), 54(3), 205-216.



## Abstract

### Objectives

To investigate the definitions and applications of 'spaced learning' and to propose future directions for advancing its study and practice in health professions education.

### Method

The authors searched five online databases for articles published on spaced learning in health professions education prior to February 2018. Two researchers independently screened articles for eligibility with set inclusion criteria. They extracted and analysed key data using both quantitative and qualitative methods.

### Results

Of the 2972 records retrieved, 120 articles were included in the review. More than 90% of these articles were published in the last 10 years. The definition of spaced learning varied widely and was often not theoretically grounded. Spaced learning was applied in distinct contexts, including online learning, simulation training and classroom settings. There was a large variety of spacing formats, ranging from dispersion of information or practice on a single day, to intervals lasting several months. Generally, spaced learning was implemented in practice or testing phases and rarely during teaching.

### Conclusions

Spaced learning is infrequently and poorly defined in the health professions education literature. We propose a comprehensive definition of spaced learning and emphasise that detailed descriptions of spacing formats are needed in future research to facilitate the operationalisation of spaced learning research and practice in health professions education.

## Introduction

The spacing effect is one of the most robust phenomena in the science of learning. Hundreds of published reports have replicated the spacing effect, originally uncovered by Ebbinghaus, which suggests that knowledge retention is enhanced when learning sessions are spaced (Cepeda et al., 2009; Carpenter et al., 2012). Re-exposing learners to information over time using temporal intervals, i.e. spaced learning, results in more effective storage of information than if it was all provided at a single time, i.e. massed learning. There is mounting evidence that students do not remember what is learned, also in health professions education (HPE) (Schneid et al., 2018; Weggemans et al., 2017; Simanton et al., 2012; Sullivan et al., 2013; Marcel, 2006). Researchers have therefore indicated a need to invest time and resources in helping learners retain the information being learned (Marcel, 2006). Educational principles grounded in a spaced learning approach have the potential to address this growing challenge in HPE.

Although literature reviews on effective learning in HPE exist and suggest a key role for spaced learning in optimising retention, systematic analysis of spaced learning research is complicated by the great diversity in the terms and definitions used in this literature, including 'distributed practice', 'spaced education', and 'retrieval practice' (Yeh & Park, 2015; Augustin, 2014; Friedlander et al., 2011; Phillips et al., 2019; van Hoof & Doyle, 2018) The variety of learning and assessment methods that are referred to as spaced learning further complicate the analysis of its effects. According to definitions used by psychologists, spaced learning should include learning sessions that are spaced over time and include repeated information (Cepeda et al., 2008). Both cumulative testing and simulation training as performed in HPE, for instance, can be considered applications of spaced learning. In addition to the variety of educational activities, spacing formats often differ in terms of their temporality, with some researchers distributing learning sessions over a few days, whereas others use hours, weeks or months. Moreover, it is often unclear if researchers used evidence from empirical research or relied on a theoretical framework to inform their spacing format. Overall, the broad range of terms associated with spaced learning, the multiple definitions and variety of applications used in HPE can hinder the operationalisation of spaced learning.

A comprehensive synthesis of the various definitions and applications of spaced learning in HPE may help identify gaps in knowledge, highlight areas for future research and support a more effective implementation of spaced learning in the HPE curricula. Therefore, the purpose of this paper was to investigate how spaced learning is defined and applied across HPE contexts.

## Methods

We employed a scoping review methodology to examine the definitions and applications of spaced learning in HPE. To execute the review in a rigorous manner, we assembled a research team consisting of co-investigators with in-depth knowledge of HPE (MV, RH, AT, BO and PS), methodological experience (AT and BO), and medical library expertise (CP). We used the methodological framework developed by Arksey and O'Malley (2005), which was later refined by Levac and colleagues (2010). The framework consists of the following six steps: Step 1, identifying the research question; Step 2, identifying relevant articles; Step 3, selecting articles; Step 4, charting the data; Step 5 collating, summarising and reporting the results, and Step 6, consultation. Step 6, consultation, was not conducted as we aimed to study the HPE literature specifically without including additional stakeholders' perspectives on this matter.

### Identifying the research question

Given our goal of identifying key concepts, and applications of spaced learning, we generated a main research question that allows for a broad exploration of spaced learning. The overarching question guiding this scoping review was as follows: 'How is spaced learning defined and applied in HPE?' Accordingly, we sought to answer the following specific research questions: (RQ1A) Which concepts are used to define spaced learning and associated terms? (RQ1B) To what extent do these terms show conceptual overlap? (RQ2) Which theoretical frameworks are used to frame spaced learning? (RQ3) Which spacing formats are utilised in spaced learning research?

### Identifying relevant studies

An university affiliated librarian (CP) was consulted when drafting the search query. An initial brainstorming session with the research team and librarian led to the inclusion of 'spaced learning' and possible associated terms, such as 'spaced training', 'spaced education', 'distributed practice', 'test-enhanced learning', and 'retrieval practice'. The final search was conducted on 28 February 2018 using five databases: PubMed, Web of Science, Embase, Education Resources Information Center (ERIC), PsycINFO (Supplementary A). MV conducted additional forward reference searching of included review articles to identify additional articles.

### Selecting the studies

There was no restriction on year of publication; therefore, all articles published up until 28 February 2018 were screened for eligibility. To be included, articles had to: (a) focus on HPE, e.g. medicine, nursing, pharmacology, and (b) explicitly name 'spaced learning', or any associated term with a spaced study format. We excluded editorials, commentaries, conference abstracts and books, as well as non-English articles.

Two researchers (MV and RH) tested the inclusion criteria on a 10% subset of titles (Mateen et al., 2013; Thomas et al., 2017). A single calibration exercise was sufficient for the team to reach full agreement after inclusion criteria were discussed and clarified. In the abstract screening stage, RH and MV tested the inclusion criteria using a subset of papers (5%). After reaching full agreement, MV independently screened the remaining abstracts. Two additional calibration exercises were performed with RH independently screening 2.5% of abstracts ($n$ = 34) halfway and again 2.5% ($n$ = 34) at the end of the process to ensure that MV's interpretation of the inclusion criteria was consistent with the original calibration outcome. Disagreements were resolved by discussion. If the focus of the article was unclear based on the title and abstract, the full article was inspected.

### Charting the data

The data charting form was developed by MV and RH based on the units of analysis included in the research questions, e.g. definition, theoretical framework, timing of events and setting, using Microsoft® Excel 2010 (Microsoft Corp., Redmond, WA, USA). They independently extracted data from five full text articles to pilot the form. The usability of the charting form was discussed and minor modifications were made accordingly, i.e. extraction categories were added and others were removed. For instance, the 'intervention design' category from a previous version of the charting form was merged with the 'timing of events' category in the final version. The process was repeated with an additional five full text articles, followed by discussion, resulting in a final extraction form comprised of the following categories: title; author; publication year; location; terms used for spaced learning; definition by researchers; theoretical framework; population; research method; research design; report of evidence-based spacing; timing of events; topic of learning; type of knowledge; setting; basic sciences/clinical, and learning phase.

### Collating, summarising and reporting the results

*Numerical analysis*

We performed a numerical analysis to describe the study characteristics, i.e. year of publication, location, population, educational content, domain, subject, theoretical frameworks (RQ2A) and spacing formats (RQ3A) included in each paper.

*Thematic analysis*

The variety of spaced learning definitions and associated terms (RQ1A) were synthesised using a thematic analysis. Two researchers (MV, RH) generated a list of open codes from words or phrases in the definitions. Discussion between the two researchers explored relationships between open codes across definitions, which we refer to as concepts. These concepts were then analysed to generate overarching core themes. Drawing from the previously identified core themes as predetermined categories, we used a deductive approach to search for conceptual overlap amongst terms and definitions (RQ1B). Cross-checking of coding strategies and interpretation of data was performed by BO.

## Results

### Descriptive summary

The database search resulted in a total of 2972 records (Figure 1). After duplicates were removed, 2184 records remained. After applying title and abstract screening criteria, we identified 270 articles as eligible for full text review. A total of 120 articles met all criteria and were retained for the full review. Of these articles, 109 (91%) were published in the last 10 years (Supplementary B). Approximately two-thirds of all studies ($n$ = 76; 63%) were conducted in the United States, 25 in Europe (20%), eight in Canada (7%), seven in Australia (6%), two in Asia (2%) and two in South America (2%). See Supplementary C for an overview of the other study characteristics.

### Definitions of spaced learning

Besides the term 'spaced learning', we found 20 associated terms used to define this concept. Some terms were found in multiple studies but were defined differently, e.g. distributed practice, others were only defined in a single study, e.g. spaced distribution, or not defined at all, e.g. spaced retrieval practice.

There was a total of 74 definitions (for an extended overview of all definitions see Supplementary D). These definitions were analysed thematically, resulting in the identification of seven core themes: Educational activity was the most recurrent theme (64/74); followed by Structure (51/74); Timing (44/74); Content (28/74); Repetition (27/44); Learning outcomes (24/74), and Educational tool (14/74). For each core theme, large variation was found amongst definitions, which resulted

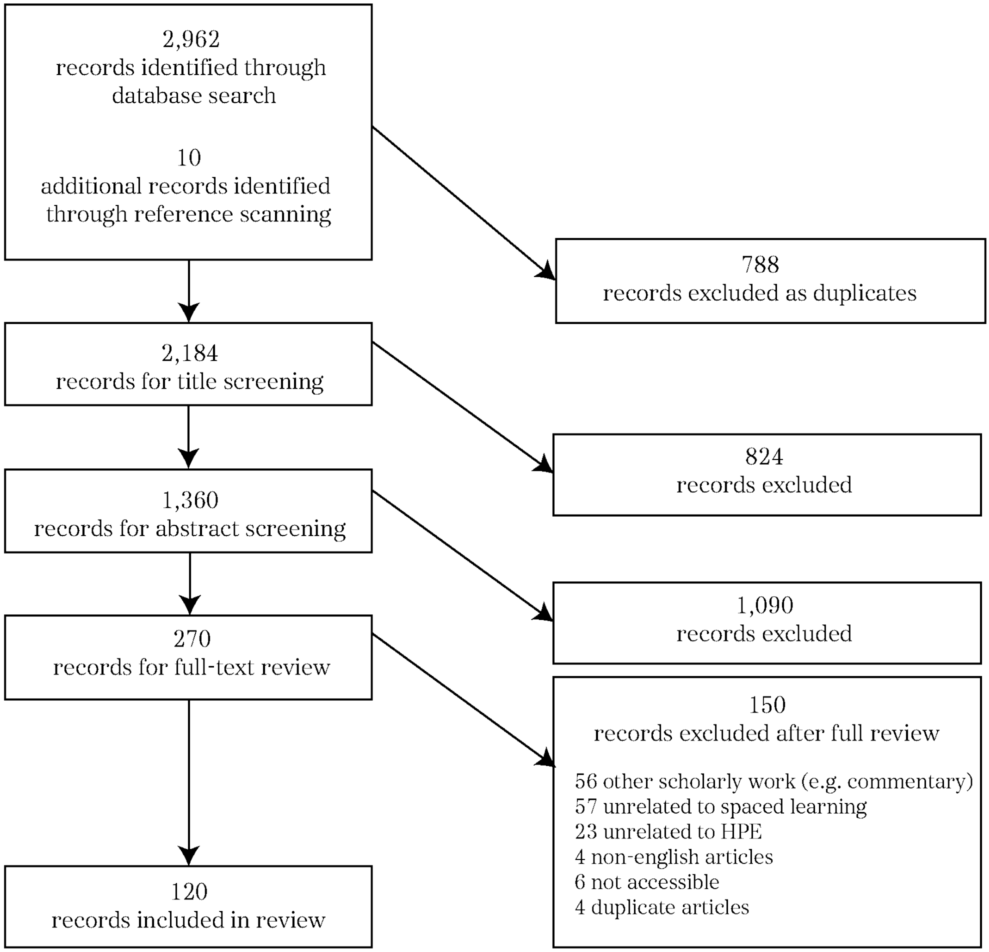



Figure 1 | Flow chart for the scoping review selection process.

in a number of sub-themes (see Table 1). For instance, an 'educational activity' was described in terms of what it should entail, e.g. listening and rereading, practicing, or what it should not entail, e.g. not highlighting, not summarising and not cramming. Additionally, some definitions encompassed specific details about the number of educational activities and the size or the division of labour.

Due to this large variation in definitions, a deductive approach was necessary to study conceptual overlap between terms and this approach was conducted on the core theme level. The recurrent core themes for each of the 21 terms are shown in Table 2. For instance, for the term 'spaced learning' we found five definitions all of which included a notion of a certain educational activity, structure and timing.

Table 1 | A thematic analysis of definitions of spaced learning and its related terms.

| First level theme (core theme) | Second level theme | Third level theme | Fourth level theme |
|---|---|---|---|
| Educational activity | Number | Singular | |
| | | Plural | |
| | Type | Reviewing | |
| | | Reading | |
| | | Test | Short |
| | | | MCQ |
| | | | Mastery |
| | | | Physical |
| | | | Not achievement |
| | | Distractor | |
| | | Listening | |
| | | Relearning | |
| | | Case-based | |
| | | Receiving feedback | |
| | | Practicing | |
| | | Studying | |
| | | Learning | |
| | | Recalling | |
| | | Not rereading | |
| | | Not relistening | |
| | | Not highlighting | |
| | | Not summarizing | |
| | | Not cramming | |
| | Size | Curricula | |
| | | Smaller | |
| | Division of labour | Providing | |
| | | Strategy for learning *(student-directed)* | |
| Structure | Dispersion | | |
| | Alternation | Irregular | |
| | | Large | |
| | Interruption of activity | Rest | |
| | | 10-20 minutes | |
| | Not packed together | | |
| | Not a single time | | |

| | | | |
|---|---|---|---|
| | No dispersion | | |
| | Adaptive | | |
| Content | Information / Content | Multiple sets | |
| | | Small | |
| | | Identical | |
| | | New | |
| | Stimuli | | |
| Repetition | Rehearsal | 3 times | |
| | | Periodically | |
| Timing | Comparative | Longer | |
| | | Later | |
| | Adjective | Long | |
| | | Short | |
| | | Fast | |
| | Specific duration | Fixed | |
| | | Days | |
| | | Weeks | |
| | | Months | |
| | | Less than 5 min | |
| | Other | Increasing | |
| | | Previous | |
| | | Immediately prior | |
| | | Concurrently with an activity | |
| Educational tool | Multi-source | | |
| | Owned by student | | |
| | Electronic | | |
| | Online | | |
| | Gamifications | | |
| Learning outcomes | Knowledge | | |
| | Skill | | |
| | Impact on behaviour | | |
| | Forgetting | Natural | |
| | Recall/remember | Retention | |
| | | Silent | |
| | Effect | More effective | Deliberate |
| | | | Adaptive |
| | | Reducing | |

Table 2 | Overview of the identified core themes in the definitions. In case of terms with plural definitions only core themes recurrent in all definitions are indicated. The number of plural definitions is indicated for each term in brackets. This table only shows the terms with definitions (21 total). Terms identified without definitions were: Dispersed learning, Distributed learning, Repeated retrieval practice, Spaced instruction, Spaced training, Spaced retrieval practice, Spaced studying, and Spaced testing.

| | | Educational activity | Structure | Timing | Repetition | Content | Educational tool | Learning outcomes |
|---|---|---|---|---|---|---|---|---|
| Plural definitions | Spaced learning (5) | 5/5 | 5/5 | 5/5 | 4/5 | 3/5 | 1/5 | 2/5 |
| | Spaced practice (2) | 2/2 | 2/2 | 2/2 | 1/2 | 1/2 | | 1/2 |
| | Retrieval practice (6) | 6/6 | 1/6 | 3/6 | | 1/6 | | 2/6 |
| | Distributed practice (19) | 19/19 | 17/19 | 14/19 | 2/19 | 5/19 | | 4/19 |
| | Spaced education (19) | 15/19 | 15/19 | 8/19 | 9/19 | 8/19 | 11/19 | 7/19 |
| Single definition | Spaced approach | X | X | X | | | | |
| | Distributed training | X | X | X | | | | |
| | Spaced distribution | X | X | X | X | | | |
| | Distributed study | X | X | X | X | X | | X |
| | Spaced repetition | X | X | X | X | X | | X |
| | Automated spaced repetition | X | X | X | X | X | | X |
| | Repeated practice | X | X | | | | | |
| | Structured spaced training | X | X | | | | | |
| | Interleaved practice | X | X | | | X | | |
| | Spaced training | X | X | | | | | X |
| | Interactive spaced education | X | X | | X | X | X | X |
| | Space repetition learning | X | | X | | X | | X |
| | Interval learning | X | | X | X | X | | |
| | Interval training | X | X | X | X | | | |
| | Repeated testing | X | | X | X | X | | X |
| | Distributed method of learning | X | | X | X | X | X | X |

### Framing spaced learning

Almost half of the empirical research articles (*n* = 48, 47%) did not explicitly mention a theoretical framework. In total, nine theoretical frameworks were mentioned in the remaining studies of which the Spacing effect[1] (*n* = 40) and Testing effect[2] (*n* = 31) were named most often. Other frameworks were Cognitive Load Theory[3] (*n* = 4), Desirable Difficulties Theory[4] (*n* = 2), Retrieval hypothesis[5] (*n* = 2), Total-time hypothesis[6] (*n* = 2), Learning Theory[7] (*n* = 1), Metacognitive Theory[8] (*n* = 1) and Kolb's Experiential Learning Theory[9] (*n* = 1).

Only a few studies[10] (*n* = 15, 15%) based their spacing format on previous empirical research. Articles by Cepeda and colleagues (2008)[11] (*n* = 7) and Pashler and colleagues (2007)[12] (*n* = 7), both derived from psychological literature on the spacing effect, were cited most often.

---

[1] Al Rawi et al., 2015; Boespflug et al., 2015; Bruckel et al., 2016; Dobson, 2011; Flett, 2017; Gandhi et al., 2016; Gyorki et al., 2013; Hernick, 2015; Horn & Hernick, 2015; Kerfoot, 2010; Kerfoot, 2009; Kerfoot, 2008; Kerfoot et al., 2008; Kerfoot & Baker, 2012a; Kerfoot & Baker, 2012b; Kerfoot et al., 2012; Kerfoot et al., 2007a; Kerfoot & Brotschi, 2009; Kerfoot et al., 2007b; Kerfoot et al., 2010a; Kerfoot et al., 2009; Kerfoot et al., 2010b; Kerfoot et al., 2011; Kerfoot et al., 2014; Mathes et al., 2014; Matos et al., 2017; Matzie et al., 2009; Miller et al., 2016; Nakata et al., 2017; Nkenke et al., 2012; Ojha et al., 2014; Patocka et al., 2015; Raman et al., 2010; Shaikh et al., 2017; Shaw et al., 2011; Shaw et al., 2012; Shenoi et al., 2016; Smeds et al., 2016; Spruit et al., 2015; Taveira-Gomes et al., 2014; Bekkink et al., 2012.

[2] Boespflug et al., 2015; Bruckel et al., 2016; Dobson, 2011; Flett et al., 2017; Gandhi et al., 2016; Gyorki et al., 2013; Hernick, 2015; Horn & Hernick, 2015; Kerfoot, 2010; Kerfoot, 2008; Kerfoot et al., 2008; Kerfoot & Baker, 2012a; Kerfoot & Baker, 2012b; Kerfoot et al., 2012; Kerfoot et al., 2010a; Kerfoot et al., 2010b; Kerfoot et al., 2014; Mathes et al., 2014; matos et al., 2017; Shaikh et al., 2017; Shaw et al., 2011; Shaw et al., 2012; Smeds et al., 2016; Bekkink et al., 2012; Burdo & O'Dwyer, 2015; Dobson et al., 2017; Freda & Lipp, 2016; Galvagno & Segal, 2009; Jackson et al., 2011; Kerfoot et al., 2009; Larsen et al., 2013a; Terenyi et al., 2018.

[3] Raman et al., 2010; Taveira-Gomes et al., 2014; Andersen et al., 2016a; Andersen et al., 2016b.

[4] Burdo & O'Dwyer, 2015; Dobson, 2012.

[5] Baghdady et al., 2014; Spreckelsen & Juenger, 2017.

[6] Baghdady et al., 2014; Spreckelsen & Juenger, 2017.

[7] Breckwoldt et al., 2016.

[8] Bude et al., 2011.

[9] Freda & Lipp et al., 2016.

[10] Horn & Hernick, 2015; Kerfoot & Baker, 2012a; Kerfoot & Baker, 2012b; Kerfoot et al., 2012; Kerfoot et al., 2007a; Kerfoot et al., 2010a; Kerfoot et al., 2011; Kerfoot et al., 2014; Spruit et al., 2015; Dobson et al., 2017; Kerfoot et al., 2009; Dobson, 2012; Scales et al., 2016; Boettcher et al., 2018; Schoeff et al., 2017.

[11] Horn & Hernick, 2015; Kerfoot & Baker, 2012a; Kerfoot & Baker, 2012b; Kerfoot et al., 2012; Kerfoot et al., 2010b; Kerfoot et al., 2014; Scales et al., 2016.

[12] Horn & Hernick, 2015; Kerfoot & Baker, 2012a; Kerfoot & Baker, 2012b; Kerfoot et al., 2012; Kerfoot et al., 2010a; Kerfoot et al., 2011; Kerfoot et al., 2014.

[13] Boespflug et al., 2015; Bruckel et al., 2016; Dobson, 2011; Gandhi et al., 2016; Kerfoot, 2010; Kerfoot, 2009, Kerfoot, 2008; Kerfoot & Baker, 2012a; Kerfoot & Baker 2012b; Kerfoot et al., 2012; Kerfoot et al., 2007a; Kerfoot & Brotschi, 2009; Kerfoot et al., 2007b; Kerfoot et al., 2010a; Kerfoot et al., 2009; Kerfoot et al., 2010b; Kerfoot et al., 2011; Kerfoot et al., 2014; Mathes et al., 2014; Matos et al., 2017; Matzie et al., 2009; Miller et al., 2016; Nkenke et al., 2012; Shaikh et al., 2017; Shaw et al., 2011; Shaw et al., 2012; Shenoi et al., 2016; Smeds et al., 2016; Kerfoot et al., 2009; Scales et al., 2016; Barsoumian & Yun, 2018; Blazek et al., 2016; Larsen et al., 2015; Pernar et al., 2013; Phillips et al., 2017; Phillips et al., 2014; Robinson et al., 2017; Tshibwabwa et al., 2017.

### Applying spaced learning

Approximately half of the empirical research articles (*n* = 51, 48%) applied spaced learning in an online setting, mostly through delivering learning sessions in e-mails distributed over time using electronic modules, e.g. Qstream[13] (*n* = 38, 37%). Spaced learning was also implemented in simulation settings[14] (*n* = 24, 23%), generally used to disperse training sessions over time to stimulate clinical skill acquisition. In total 24 studies[15] (23%) were conducted in classrooms and applied to various educational activities, ranging from repeated practice and testing of basic science mechanisms, to clinical scenarios and skill training.

The spacing formats of experimental and observational studies were analysed and summarised for the three different settings that were identified previously, that is online, simulation and classroom settings.

For the online setting, the duration of events showed a great variety between studies. Information or questions were distributed through online sources daily[16] (*n* = 10), every 2 days[17] (*n* = 8), every 3 days[18] (*n* = 1), weekly[19] (*n* = 12), every 2 weeks[20] (*n* = 1), or monthly[21] (*n* = 2). In studies explicitly stating that material was

[14] Nakata et al., 2017; Ojha et al., 2014; Shaw, 2012; Spruit et al., 2015; Larsen et al., 2013a; Andersen et al., 2016a, Andersen et al., 2016b; Boettcher et al., 2018; Schoeff et al., 2017; Akdemir et al., 2014; Andersen et al., 2015; Bjerrum et al., 2016; Connor et al., 2016; Ernst et al., 2014; Fann et al., 2008; Mackay et al., 2002; Moulton et al., 2006; Nesbitt et al., 2013; Stransky et al., 2010; Stefanidis et al., 2006; Kesser et al., 2014; Verdaasdonk et al., 2007; Mitchell et al., 2011; Kurosawa et al., 2014; Stefanidis et al., 2009.

[15] Patocka et al., 2015; Raman et al., 2010; Bekkink et al., 2012; Burdo & O'Dwyer, 2015; Dobson et al., 2017; Freda & Lipp, 2016; Larsen et al., 2013a; Terenyi et al., 2018; Baghdady et al., 2014; Breckwoldt et al., 2016; Bude et al., 2011; Stransky et al., 2010; Kleiman et al., 2017; Moore & Chalk, 2012; Murrihy et al., 2009; Rozenshtein et al., 2016; Barrington et al., 2016; Ayyub & Mahboob, 2017; Akkaraju, 2016; Dobson, 2013; LaBossiere et al., 2016; Larsen et al., 2013b; Halliday et al., 2015; Marcotte et al., 1976.

[16] Flett et al., 2017; Kerfoot, 2010; Kerfoot, 2009; Kerfoot, 2008; Kerfoot et al., 2008; Kerfoot et al., 2012; Kerfoot et al., 2007a; Kerfoot et al., 2010a; Smeds et al., 2016; Tshibwabwa et al., 2017.

[17] Boespflug et al., 2015; Gandhi et al., 2016; Gyorki et al., 2013;,Kerfoot & Baker, 2012a; Shaw et al., 2011; Shaw et al., 2012; Kerfoot et al., 2009; Barsoumian & Yun, 2018.

[18] Kerfoot et al., 2014.

[19] Bruckel et al., 2016; Kerfoot & Brotschi, 2009; Kerfoot et al., 2007a; Kerfoot et al., 2010a; Kerfoot et al., 2011; Matos et al., 2017; Matzie et al., 2009; Spreckelsen & Juenger, 2017; Larsen et al., 2015; Pernar et al., 2012; Raupach et al., 2016; Taveira-Gomes et al., 2015.

[20] Miller et al., 2016.

[21] Matos et al., 2017; Blazek et al., 2016.

[22] Flett et al., 2017; Smeds et al., 2016.

[23] Boespflug et al., 2015; Gandhi et al., 2016; Gyorki et al., 2013; Kerfoot, 2010; Kerfoot, 2009, Kerfoot, 2008; Kerfoot & Baker, 2012a; Kerfoot & Baker, 2012b; Kerfoot et al., 2012; Kerfoot et al., 2007a; Kerfoot & Brotschi, 2009; Kerfoot et al., 2010a; Kerfoot et al., 2011; Kerfoot et al., 2014; Mathes et al., 2014; Matos et al., 2017; Shaikh et al, 2017; Shaw et al., 2011; Shaw et al., 2012; Scales, 2016; Barsoumian & Yun, 2018; Larsen et al., 2015; Robinson et al., 2017; Tshibwabwa et al., 2017.

[24] Boespflug et al., 2015; Kerfoot et al., 2008; Kerfoot & Baker, 2012b; Kerfoot et al, 2012; Kerfoot et al., 2007b; Kerfoot et al., 2010a; Kerfoot et al., 2010b; Kerfoot et al., 2011; Matos et al., 2017; Matzie et al., 2009; Miller et al., 2016; Shenoi et al., 2016; Kerfoot et al., 2009.

[25] Spruit et al., 2015; Boettcher et al., 2018; Bjerrum et al., 2016; Mackay et al., 2002; Stransky et al., 2010; Kurosawa et al., 2014; Barrington et al., 2016.

not only spaced but also repeated, repetition delays ranged from various days[22] ($n$ = 2), to weeks[23] ($n$ = 24), to months[24] ($n$ = 12). Additionally, there were large variations in the number of repetitions and intervals between repetitions.

For the simulation setting, studies frequently used designs in which training sessions were distributed within a single day[25] ($n$ = 7) or within a set number of consecutive days, weeks or months[26] ($n$ = 15). Notably, there were numerous differences in the number of training sessions, total training time and duration of intervals.

For the classroom setting, most studies described the use of interim, e.g. cumulative testing[27] ($n$ = 15) to enhance long-term retention of to-be-learned information. Other applications of spaced learning in the classroom involved the distribution of teaching or learning sessions over multiple days[28] ($n$ = 3), weeks[29] ($n$ = 3), or months[30] ($n$ = 1). It was often unclear if sessions included repetition of material taught during preceding sessions or if each session solely consisted of new material.

Studies were mainly concerned with improving the effectiveness of learning through spacing of practice and/or testing ($n$ = 91, 88%). Only four studies[31] (4%) focused efforts on spaced learning as a means of teaching, for example, during conventional lectures.

## Discussion

We conducted a scoping review to examine how spaced learning is defined and applied in HPE. Spaced learning appeared relatively new to HPE, with 90% of the articles in our review having been published only in the last 10 years. This is an interesting finding given that the first description of the spacing effect dates back to 1885 and has been a major subject of research in the educational psychology literature since (Ebbinghaus, 1885). Our findings indicate that most spaced learning applications in HPE involve online learning, which may explain the later presence of spaced learning in our field. In light of the increasing popularity of spaced learning in HPE, it is concerning that descriptions of its applications lack the necessary detail to support implementation or replication. Our review showed

[26] Spruit et al., 2015; Andersen et al., 2016a; Andersen et al., 2016b; Schoeff et al., 2017; Akdemir et al., 2014; Andersen et al., 2015; Bjerrum et al., 2016; Moulton et al., 2006; Stransky et al., 2010; Stefanidis et al., 2006; Kesser et al., 2014; Verdaasdonk et al., 2007; Mitchell et al., 2011; Kurosawa et al., 2014; Barrington et al., 2016.

[27] Bekkink et al., 2012; Freda & Lipp, 2016; Galvagno & Segal, 2009; Larsen et al., 2013a; Terenyi et al., 2018; Baghdady et al., 2014; Moore & Chalk, 2012; Ayyub & Mahboob, 2017; Akkaraju, 2016; Dobson, 2013; LaBossiere et al., 2016; Larsen et al., 2013b; Kerkdijk et al., 2015; Kerdijk et al., 2013; LaDisa & Biesboer, 2017.

[28] Dobson et al., 2017; Breckwoldt et al., 2016; Halliday et al., 2015.

[29] Patocka et al., 2015; Raman et al., 2010; Burdo & O'Dwyer, 2015; Bude et al., 2011.

[30] Bude et al., 2011.

[31] Raman et al., 2010; Pernar et al., 2013; Marcotte et al., 1976; Pernar et al., 2012.

that in most research spaced learning is poorly defined and almost half of the studies do not explicitly mention a theoretical framework. Even fewer studies based their spacing formats on empirical literature. It is possible that these shortcomings may be linked to the presence of 'innovators' and 'early adopters' in our field. According to Rogers' Diffusion of Innovation Theory (Rogers, 2003),

these groups value the trialability attribute of innovations, i.e. how easily potential adopters can explore your innovation, which aligns with our findings. All spaced learning studies in HPE that we analysed were conducted in authentic educational environments instead of laboratory settings. As such, the focus may be on improving educational practices and less on advancing theory or knowledge. However, this approach makes replication and follow-up of current studies on spaced learning challenging. Clearer definitions and detailed descriptions of applications are needed for scholars and educators to improve future research and practice on spaced learning in HPE.

**Defining spaced learning**

We examined 74 definitions of spaced learning and associated terms. Concepts found amongst these definitions were organised into seven core themes: Educational activity; Structure; Timing; Repetition; Educational tool, and Learning outcomes. Most terms were defined by unique combinations of core themes resulting in low conceptual overlap between terms. Additionally, some terms seemed to relate to a more specified version of spaced learning as hey contained more core themes than others. For instance, the definition of 'spaced repetition' includes the notion of 'reviewing of content multiple times over optimised time intervals', whereas 'spaced approach' limits itself to stating 'the distribution of fixed teaching hours over a longer time period.' It is important to note that the core themes were derived from a large variety of second to fourth level themes, illustrating the vagueness of definitions. For example, the educational activity as mentioned in the definition of 'spaced distribution', concerns the number of activities, whereas a definition of 'spaced learning' focuses on the type of activities, i.e. tests. Although they both say something about learning engagement, they differ in what information they deem relevant.

Furthermore, different definitions of the same term typically showed few recurrent core themes suggesting low conceptual overlap. For example, we found that the five definitions of the term 'spaced learning' shared the following core themes: Educational activity; Structure, and Timing; whereas Education tool was only found in one of the definitions.

Clearly, there is no unified definition of spaced learning in the HPE literature.

We suggest that a more consistent use of terminology can facilitate a more systematic appraisal of future research. Based on our findings we propose the following comprehensive definition of spaced learning, which explicitly covers all involved components:

*Spaced learning involves [specified] educational encounters that are devoted to the same [specified] material, and distributed over a [specified] number of periods separated by a [specified] interstudy interval, with a [specified] learning outcome after a [specified] retention interval.*

These components should be clearly specified for each study on spaced learning to facilitate comparison and crosstalk between spaced learning researchers in our community.

**Framing spaced learning**

There is room for improvement regarding framing of the spaced learning concept as almost half of the articles did not explicitly frame their research using a theoretical framework. This might be related to the diversity and vagueness amongst terms used to define spaced learning, which may have complicated researchers' search for previous empirical research and associated theoretical frameworks. These findings are illustrative of the general underuse of theory in HPE research (Bordage, 2009; Laksov, 2017). Importantly, use of theory can help educators and researchers to better understand existing problems and formulate new research questions.

**Applying spaced learning**

Spaced learning is applied broadly in HPE, spanning various health professions, subjects, and educational settings, i.e. online, simulation and classroom. Exploring the specific details of its applications was rather challenging due to the absence of vital information on used spacing formats such as the number and duration of intervals between educational encounters, the duration of the retention interval, and the number and duration of learning sessions. We emphasise that in future research, spacing formats should be reported in detail to ensure reproducibility and generalisability of the outcomes (Young et al., 2018; Shea et al., 2004).

During educational encounters, spacing formats mostly included spaced learning in the testing or practice phase. The occurrence of the 'testing effect' as the second most used theoretical framework fits this application pattern. Notably, less research is conducted on the benefits of spaced learning in the instructional phase, that is during teaching. We consider this a gap in the literature and propose that HPE may draw from the rich scientific literature on spaced learning in education and psychology to develop spaced learning formats that can optimise the retention of knowledge. Psychological and neuroscientific research findings on the mechanisms of memory formation suggest that spaced learning also works using shorter intervals (Horn & Hernick, 2015). Therefore, applying spaced learning on the timescale of minutes to hours may have implications for current massed learning in classroom settings, such as conventional lectures, which still holds a prominent position in HPE worldwide. Ultimately, implementing and optimising spaced learning formats across curricula may help to prepare health professionals with a solid foundational body of knowledge.

### Limitations

Although we attempted to be as thorough as possible, our search was limited to the selected databases, search terms and English-written scholarly articles, which may have excluded relevant articles inadvertently. Furthermore, as a scoping review aims to investigate the nature and extent of the research topic, we did not critically appraise the included studies.

## Conclusion

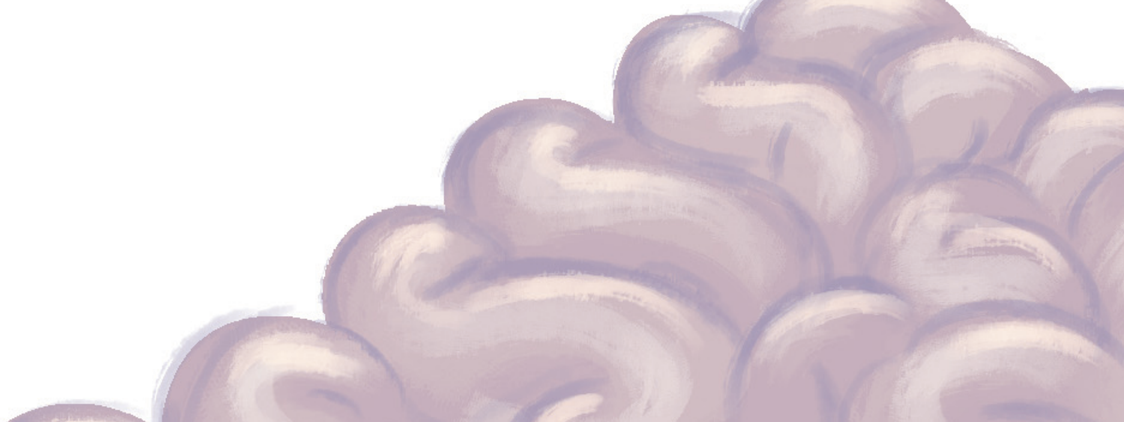
This scoping review has highlighted the large variety in definitions and applications of spaced learning across HPE. Based on our findings and our review of the psychological and neuroscientific literature, we offer the following recommendations to improve research and educational practice related to spaced learning: (a) define the spaced learning concept in an explicit and comprehensive manner in order to stimulate consistent application; (b) use study designs that are described thoroughly and informed by empirical research on spaced learning, related theories, and practices, and (c) further expand the spaced learning applications beyond online learning and simulation training, for example, by applying spaced learning in the instructional phase. With these recommendations, we aim to promote an enriched understanding of spaced learning and support the development of optimal spaced learning environments in HPE curricula.

# Chapter 3

# Making a lecture stick: the effect of spaced instruction on knowledge retention in medical education

Marnix C. J. Timmer · Paul Steendijk · Sandra M. Arend · Marjolein Versteeg

Published in: Medical Science Educator (2020), 30, 1211-1219.



## Abstract

### Introduction

Poor knowledge retention is a persistent problem among medical students. This challenging issue may be addressed by optimising frequently used instructional designs, such as lectures. Guided by neuroscientific literature we designed a spaced learning lecture in which the educator repeats the to-be-learned information using short temporal intervals. We investigated if this modified instructional design could enhance students' retention.

### Materials and methods

Second-year medical students ($N$ = 149) were randomly allocated to either the spaced lecture or the traditional lecture. The spaced lecture consisted of three 15-minute instructional periods, separated by 5-minute intervals. A short summary of the preceding information was provided after each interval. The traditional lecture encompassed the exact same information including the summary in the massed format, thus without the intervals. All students performed a baseline knowledge test two weeks prior to the lectures and students' knowledge retention was assessed eight days after the lectures.

### Results

The average score on the retention test ($\alpha$ =.74) was not significantly different between the spaced lecture group (33.8±13.6%) and the traditional lecture group (31.8±12.9%) after controlling for students' baseline-test performance ($F(1,104)$ = 0.566, $p$ = .458). Students' narrative comments showed that the spaced lecture format was well received, and subjectively benefitted their attention span and cognitive engagement.

### Discussion and Conclusion

We were unable to show increased knowledge retention after the spaced lecture compared to the traditional lecture. Based on these findings, we provide recommendations for further research. Ultimately, we aim for optimised spaced learning designs to facilitate learning in the medical curriculum and to help educate health professionals with a solid knowledge base.

## Introduction

Medical students have a hard time recalling knowledge they acquire during medical training (Weggemans et al., 2017; Schneid et al., 2018; Simanton et al., 2012; Custers & ten Cate, 2011). Since successful clinical reasoning is built upon a solid foundation of knowledge, medical education is facing a serious problem (Herwaarden et al., 2009). This issue of forgetfulness may partly result from currently used teaching practices Namely, lectures in which a large volume of information is covered over a short uninterrupted time-span, so called massed learning, are still commonly used as a teaching modality by medical educators, especially in basic sciences education. This approach has shown to be rather ineffective when aiming for long-term knowledge retention (Rawson & Kintsch, 2005). One may consider adjusting such practices by using an alternative strategy, i.e. repeating information in several learning sessions distributed over time. This spaced learning approach is based on results from a century of psychological research (for a meta-analysis see Cepeda et al., 2006 or Carpenter et al., 2012) and could be a valuable addition to instructional designs in medical education.

The spacing effect is a robust phenomenon that forms the basis of spaced learning methods. During spaced learning, knowledge or skills that have to be acquired are repeated in several learning sessions that are distributed over time. Spaced learning is usually contrasted with massed learning where information is packed together in a single learning session and only repeated consecutively, if repeated at all. The beneficial effects of spaced learning on retention have been shown for a variety of learning tasks concerning factual knowledge, e.g. Cepeda et al., (2008), conceptual knowledge, e.g. Gluckman et al., (2014), Rohrer & Taylor, (2007), and procedural knowledge, e.g. Simmons, (2011).

Over the last ten years, research has proven spaced learning to be successful in various medical disciplines, including surgery, urology, radiology and general clinical reasoning (Gyorki et al., 2013; Kerfoot et al., 2007a; Smeds et al., 2016; Nkenke et al., 2012; Boettcher et al., 2018; Moulton et al., 2006; Patocka et al., 2015). Spaced learning has mostly been investigated in online learning environments and simulation settings. In surgical skill training, for example, a recent systematic review showed that spacing practice sessions resulted in increased retention of skills compared to massed training (Cecilio-Fernandes et al., 2018). Similar results have been found for spacing instructional designs. For instance, a study has shown that the dispersion of 4hrs of direct instruction over 4 weeks, i.e. 1hr/week, significantly enhanced knowledge retention after one month (Raman et al., 2010). Interestingly, neuroscientific research on mechanisms of memory implies that the spacing effect may already occur using much shorter intervals in the timescale of minutes to hours (for a review, see Smolen et al., 2016). This notion gave rise to our idea of implementing spaces within traditional massed 45-60 minute lectures to promote long-term knowledge storage among medical students.

Researchers in higher education have already reported the successful application of spaces with short intervals. Kelley and Whatson (2013) compared a 4-month biology course with a single 60-minute spaced learning session. Students following the spaced learning session repeatedly received an intensive 20-minute presentation (three times in total), intervened by 10-minute breaks. During the breaks they were asked to perform physical distractor activities, e.g. to clay, take a walk, or play basketball. These physical tasks were specifically selected to prevent any interference with the memory formation process regarding the learning material. Students' final test results were compared between groups per hour of education, and outcomes highly favoured the spaced learning cohort. This study showed that implementation of spaced learning in an instructional setting can establish long-term memory rapidly. Recent initiatives were inspired by these findings and illustrated the benefits of spaced instruction in different educational contexts (O'Hare et al., 2017; Garzia et al., 2016).

Educational initiatives promoting the use of short intervals to enhance long-term memory formation are inspired by neuroscientific evidence regarding the mechanisms of memory. An important phase in the process of long-term memory formation is the stabilization of a memory trace after the initial acquisition, referred to as consolidation (Kandel et al., 2014). Research has shown that consolidation of a memory on the molecular level, referred to as long-term potentiation (LTP), is elicited particularly by spaced trials and to a lesser extent by massed trials (Cao et al., 2014; Mauelshagen et al., 1998). A hypothesis has been formulated stating that adequately spaced stimuli overcome a refractory period that is needed for reinforcement of LTP (Smolen et al., 2016). In more detail, this refractory period may provide neurons time that is needed to synthesise molecular factors and/or facilitate feedback loops that underlie the initiation of LTP. In line with this reasoning, massed stimuli would not produce sufficient levels of molecular factors needed to support LTP even if the stimuli are repeated. Another hypothesis states that separate rounds of stimuli induce 'priming': meaning that LTP can be formed by a first stimulus, and is strengthened by a properly timed second stimulus (Smolen et al., 2016). Both hypotheses are not mutually exclusive and congruent with consolidation theory. Importantly, memory consolidation involves various molecular processes that each have their own temporal dynamics. Further research on long-term memory encoding in the brain has suggested that some specific molecular processes underlying LTP occur on the timescale of minutes and may contribute to the superiority of spaced learning (Genoux et al., 20020; Farah et al., 2009; Pagani et al., 2009; Ajay et al., 2004; Naqib et al., 2012; Xue et al., 2011; Menzel et al., 2011; Fields, 2005).

Based on previous educational experiments in higher education and the evidence derived from the neuroscientific framework, we aimed to examine the effect of short spaces on knowledge retention in medical students. Therefore, we compared a spaced lecture design with a traditional massed lecture and measured students' knowledge retention. We believe that the potential benefits of incorporating short spaces during teaching might help medical educators to make their lectures more effective.

## Methods

### Participants and setting

Second-year medical students enrolled in a course on disease mechanisms at the Leiden University Medical Center (LUMC) were invited to voluntarily participate in the study. More than 80% of contact hours in this course consists of lectures. The intervention was conducted in a lecture on the Dutch national vaccination program. In previous academic years information about the national vaccination program was covered by a self-study assignment. This topic was selected for this spaced learning study specifically, because students had received no prior formal education on this topic. The lectures were delivered as live presentations in a lecture hall, supported by a digital slideshow (Microsoft Powerpoint). This is common practice for lecturing at the LUMC.

### Ethical considerations

Study participation was on a voluntary basis as the lectures were not mandatory. Students were notified that the supplied information was part of their exam material. Those who decided not to attend any session could still access the exam material using the existing self-study assignment. Students autonomously decided if their test results could be used for research purposes by signing the informed consent form prior to the baseline-test, and again prior to the retention-test. They were informed that data would be anonymised and that they could withdraw their consent at any given time. Moreover, they were ensured that the test results would not affect their course grades. Students did not receive any additional credit for their participation. The study protocol was reviewed and approved by the Educational Research Review Board of the LUMC: OEC/ERRB/20180612/2.

**Study design**

This was an experimental study for which two lectures were designed: an experimental lecture based on spaced learning principles, i.e. spaced lecture, and a control lecture using the traditional, massed approach, i.e. traditional lecture. Participants were randomly allocated to one of the lecture sessions. The spaced and traditional lecture were held consecutively on one day to facilitate that both lectures could be given by the same lecturer (SMA). The lectures were video recorded, to explore and reveal any substantial differences if suspected by the test results. The lecturer was a highly experienced teacher and is considered an expert in the field of infectious diseases. To assess students' knowledge at baseline, they were tested approximately two weeks prior to the intervention, i.e. baseline-test. The retention-test was taken 8 days after the intervention. For a detailed scheme of the study procedure, see Figure 1.

**Intervention**

The lectures were designed for this experiment specifically, serving as a substitute for the self-study assignment that was used during preceding academic years. The lecture material comprised characteristics of the diseases covered by the Dutch national vaccination program (Topic A), type and moment of vaccinations (Topic B) and a regional and international comparison of participation and program components (Topic C). Both lectures used the same supporting slides in identical order.

*Spaced lecture*

The total presentation-time of 45 minutes was divided over three instructional periods of approximately 15 minutes, separated by breaks, i.e. intervening gaps of 5 minutes, resulting in a 60-minute lecture. Each break was followed by a short rehearsal of the previously discussed information presented by the lecturer using 2-3 summary slides (Figure 1). We explicitly chose for summary slides as a passive rehearsal strategy, to be able for the lecturer to continue direct instruction after each break. In this way, we could study the effect of spacing on knowledge retention specifically, instead of inducing additional effects caused by active retrieval. The rationale for the 5-minute spaces was based on our interpretation of neuroscientific literature, similar educational implementation studies, and practical feasibility (Smolen et al., 2016; Kelley & Whatson, 2013; Menzel et al., 2001; Fields, 2005). Topics A, B and C were allocated to instructional periods one, two and three respectively. During the 5-minute breaks, students performed distractor activities (in our case three different origami tasks) that were not in any way related to information provided in the lecture, as this is considered to prevent possible cognitive interference with the memory formation process, conform the design of Kelley and Whatson (2013).

*Traditional lecture*

The traditional lecture followed the conventional setup for lectures in medical education at the LUMC, which is a 45-minute presentation without breaks. To control for potential confounders and ascertain same time on task, the traditional lecture also contained the summary slides, i.e. repetition, and the same amount of time dedicated to distractor activities (Figure 1). The latter were scheduled before and after the 'massed' instructional session.

**Baseline-test and Retention-test**

The baseline-test and retention-test consisted of short open-ended questions. The test questions were designed by the researchers and evaluated by two independent test-experts. The lecturer was not involved in designing these tests and was not

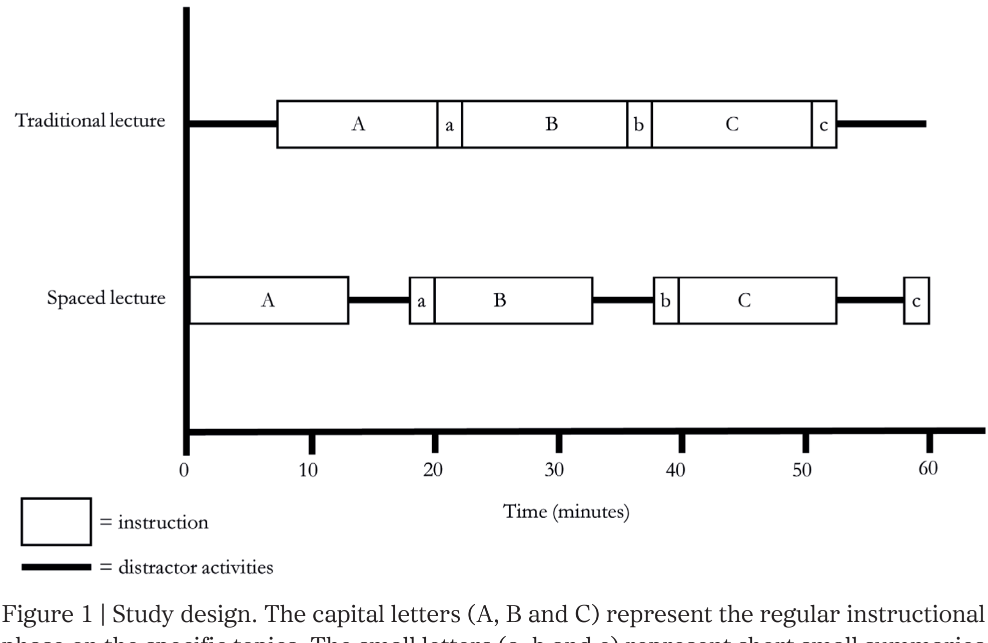



Figure 1 | Study design. The capital letters (A, B and C) represent the regular instructional phase on the specific topics. The small letters (a, b and c) represent short small summaries of the previous instruction block. In the experimental group, the regular instructional phase and small summaries were intervened by a 5-minute gap, where students were asked to perform an origami task. The traditional lecture was preceded and followed by an origami task, but the instructional phase lasted 45 consecutive minutes. Both lecture sessions ended after 60 minutes.

allowed to view the questions during the experiment. The tests were validated by performing a pilot-test with independent associates (three PhD candidates with a medical degree and two undergraduate medical students), which reassured a score of 100% could be obtained by deriving the correct answers from the lecture slides.

Participants were tested for recall of factual information that was covered by the lecture. The test consisted of short open-ended questions to reassure that recall was assessed rather than recognition. Students could obtain one point for each correct answer. If students were asked to mention two or more aspects or items in their answers to a specific question, e.g. "Which two human papilloma virus (HPV) serotypes are primarily targeted by the HPV-vaccination?", they were only awarded the point if their answer was completely correct. No penalty was given for incorrect answers. Using a pre-made answers key, students' answers were scored by one of the investigators, who was blinded for the lecture-condition. All answers considered eligible for discussion were discussed by two researchers until consensus was reached.

*Baseline-test*

The baseline-test was used to assess baseline knowledge regarding the topics covered in the lecture. The test consisted of 10 short open-ended questions. The test was performed in a lecture hall two weeks prior to the intervention to minimise priming effects.

*Retention-test*

Eight days after the intervention, students were invited to perform a retention-test. Students were requested not to study the lecture material between the lectures and the retention-test. The test included 30 short open-ended questions: the 10 questions of the baseline-test plus 20 new questions. Several additional questions were included to reveal students who had violated the study-protocol, e.g. "Did you study the lecture material between lecture and the retention-test? If yes, in what way?"

**Data collection**

All attendants at the baseline-test were linked to an anonymous study code. This study code was used to couple retention-test and baseline-test data. If students admitted they had studied or did not follow their allocated lecture, i.e. violated the study protocol, they were marked and excluded from further analyses.

### Outcome measures

The primary outcome measure was students' performance on the full retention-test. Final scores were expressed as the percentage of the maximal score. Secondary outcome measures were: i) performance on the 20 new retention-test questions and on the 10 baseline-test questions included in the retention-test, ii) narrative comments of students and the lecturer on the used lecture-formats. Test and item analyses were conducted for assessment of internal consistency and item-characteristics (Supplementary E).

Narrative comments from the students and from the lecturer were gathered for qualitative assessment. To this end, students were actively encountered by the researchers immediately following the lectures and the retention-test, and they were encouraged to express their thoughts on the lecture format in an informal way. At that time, students were unaware whether they were part of the intervention or control group. Narrative comments for both sessions were noted and stored digitally afterwards. The lecturer was interviewed after both lectures were concluded.

### Statistical analyses

For the power-analysis, the researchers agreed that a mean difference of at least one standard deviation should be detected, as this was regarded relevant for practice. Consequently, a minimum of 42 students (21 in each group) was needed to achieve 90% power at two-sided 5% significance.

The average total scores on the retention-test and the subscores for the 20 new questions and for the 10 repeated baseline-test questions were compared between study groups using ANCOVA tests, adjusting for students' baseline knowledge scores. Reliability and item characteristics (difficulty and distinctiveness) of the retention-test were evaluated by Cronbach's alpha, p-values and Rir-values (Supplementary E) (Berkel & Bax, 2006). Only data of students who completed both the baseline- and the retention-test were included for analysis. Those that did not follow their allocated lecture or restudied lecture material between the lecture and retention-test violated the study-protocol and consequently were marked and excluded from the analyses. Sensitivity analyses were carried out to reveal any major influence of these students on the outcome measures. If any of the test-questions should be removed for any cause, another sensitivity analysis for its effect on score differences would be carried out.

## Results

A total of 344 students were enrolled in the course, of whom 171 students completed the baseline-test at the start of the course (Figure 2). Halfway through the course, 149 of these students attended the lectures. Half of the participating students followed the spaced lecture (*n* = 74) and the other half followed the traditional lecture (*n* = 75). One week after the lectures, 116 students completed the retention-test. 9 participants violated the study protocol, resulting in data of 107 students to be included for final analyses. On the retention-test, there was a higher participation rate among students who attended the spaced lecture (86.5%) compared to the traditional lecture (68.0%) attendees ($\chi^2(1, 149) = 7.228$, $p < .007$). However, further statistical analyses did not need to be adjusted since groups showed similar variance for all outcome measures.

### Demographics

The mean age of participants was 19.3 ± 0.9 years (Table 1). Of these participants 90 (77.6%) were women, which resembles the overall gender distribution in the LUMC medical school. Both groups had similar average scores on the baseline-test (Spaced: 10.8% ± 8.8%, Traditional: 10.4% ± 8.8%).

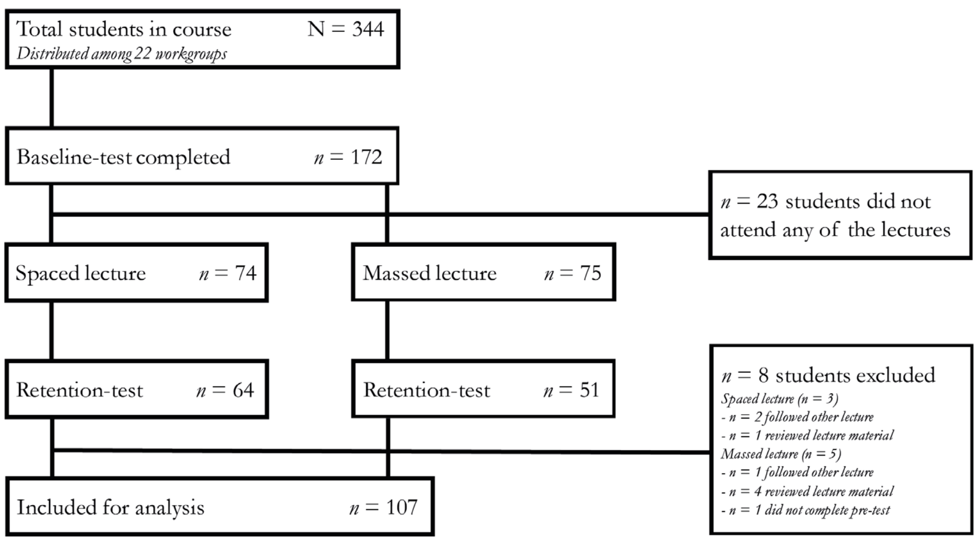



Figure 2 | Participant flow diagram.

Table 1 | Group characteristics.

| Variable | Spaced lecture (n = 61) | Traditional lecture (n = 46) |
|---|---|---|
| Gender, female | 68.9% | 89.1% |
| Age, M (SD) | 19.3 (0.9) | 19.3 (0.9) |
| Pretest performance, M (SD)[1] | 10.8% (8.8) | 10.4% (10.1) |

[1]*Maximum score was 90%, after exclusion of question 9.*

**Retention**

The spaced lecture group obtained a higher average score on the retention-test compared with the traditional lecture group (Spaced: 33.8% ± 13.6%, Traditional: 31.9% ± 12.9%), which, however, was not statistically significant ($F(1,104) = 0.566$, $p = .454$), also see Table 1. Separating the performance outcomes on the repeated original baseline-test items from the novel items revealed no significant differences between groups.

On one question of the retention-test the spaced lecture group outperformed the traditional lecture group (Spaced: 59.0% correct, Traditional: 2.2% correct). This difference likely resulted from extra information that was unintentionally provided by the lecturer during the spaced lecture. Consequently, we excluded this question in our analyses which resulted in a maximum score of 9 points on the baseline-test and 29 points on the retention-test, respectively. An alpha of 0.74 was obtained for the retention-test indicating acceptable internal consistency (see the Supplementary E for a full summary on test psychometrics).

**Sensitivity analyses**

Sensitivity analyses were conducted to investigate possible influences of the excluded data on the main results. Firstly, including participants that violated the study protocol in our analyses did not result in any significant differences versus the current reported results. Secondly, the mean difference on total retention-test scores between groups was higher when the aberrant item of the baseline-test and retention-test was not excluded, but remained not significant ($F(1,104) = 2.199$, $p = .141$). However, inclusion of this data significantly influenced the difference in retention-test scores regarding the subset of original baseline-test items ($F(1, 104) = 3.956$, $p = .049$).

### Narrative comments

Generally, students who followed the spaced lecture responded positively towards the spaced lecture format. After the retention-test one student noted: "I really hope that we [spaced lecture group] did better on the test, as I would really like to do this more often." Several participants of the spaced lecture mentioned a positive effect of the intervening gaps on their attention and productivity. For example, a student said: "During those breaks I was totally distracted from the lecture material, resulting in a feeling of total reset. I really liked this as it enhanced my attention during the whole session." Others were more doubtful: "I am not sure if the breaks improved my attention because I had a hard time to reboot after each break, therefore missing most of the small summaries." The origami task was emphasised as an enjoyable distractor activity. "Normally, everyone grabs for his or her smartphone during occasions like this [the breaks]. Now, everyone started the break trying to complete the origami task, and only switched to their phones when they gave up or failed." Another student marked the negative side effect of this particular task: "…at sudden moments I was too busy on thinking of what the next origami model would be than on the actual lecture content".

Lastly, there were some comments on the intensity of the spaced lecture: "I found it rather intense, I would hate to think of doing this four times in a row, but I could imagine it being preferable for a revision lecture."

Students who followed the traditional lecture had some positive comments on the structure of the lecture but they generally agreed that they did not notice any difference with a normal lecture. On the question: "did you experience any differences in the lecture apart from starting and ending with the origami task?",

Table 2 | Test results for spaced lecture and traditional lecture cohorts on the retention-test.

| | Spaced lecture *M (SD)* | Traditional lecture *M (SD)* | Difference (95% CI) | F | *p* | Effect size[1] ($\eta^2$) |
|---|---|---|---|---|---|---|
| Retention-test score (%) | 33.8 (13.6) | 31.8 (12.9) | 2.0 (-3.1;7.2) | 0.566 | .454 | .005 |
| Retention-test score on 20 new questions (%) | 36.6 (14.9) | 33.6 (13.2) | 3.0 (-2.5;8.5) | 1.108 | .295 | .011 |
| Retention-test score on 9 baseline-test questions (%) | 27.7 (14.7) | 27.8 (15.7) | -0.1 (-6.0;5.8) | 0.007 | .934 | .000 |

[1]*Partial $\eta^2$ effect size as predicted by the ANCOVA model*

one student answered: “hmm... well, to be honest, no.”

The lecturer enjoyed the spaced lecture format as she could recover during the breaks and experienced less fatigue afterwards. She noted: “I really had the idea that students’ attention on the last part [of the lecture] was higher than normal.”

## Discussion

This study investigated the effect of spaced learning during a lecture on medical students’ knowledge retention. We hypothesised that incorporating short spaces in the lecture would increase its effectiveness. We used an experimental design, comparing a spaced lecture with a traditional massed lecture. Our results showed that the effect of both lecture formats was not significantly different. Notably, the positive narrative comments indicated that the spaced lecture format was generally well-received by students.

In our study, we incorporated short 5-minute gaps between instruction sessions in a lecture to enhance the memory formation process. However, beneficial effects on knowledge retention were not found, suggesting that 5-minute spaces might have been too short to stimulate the consolidation process. They may have been insufficient to overcome the refractory period needed for stabilization of the memory trace, for example. Apparently, 10-minute spaces seemed to be more effective as Kelley and Whatson (2013) were successful with their spaced learning strategy in the classroom where they incorporated short 10-minute gaps. However, one should be careful interpreting these results, since findings can be highly dependent on the study design . For instance, in our study we measured knowledge retention at 8 days, whereas Kelley and Whatson measured it at 5 days. The 8-day period was selected since the Ebbinghaus forgetting curve indicates that forgetting declines exponentially and most of the forgetting occurs in the first week after initial learning (Ebbinghaus, 1885; Murre & Dros, 2015). It might be that a shorter retention period, i.e. less than eight days, had raised the ability to reveal differences between the study-groups. This is in line with evidence indicating that short intervening gaps potentially promote advantages on shorter retention periods (Cepeda et al., 2006; Cepeda et al., 2008; Cepeda et al., 2009). Another notable difference is that we chose to incorporate small summaries of preceding information in our lectures, whereas Kelley & Whatson repeated their 15-minute instructional blocks three times. Our rationale for this design was that it was closer to the traditional teaching style and was expected to be an easy-to-incorporate tool for medical educators if it was found to be effective.

Despite some empirical evidence including our own study, researchers acknowledge that optimal spacing protocols for humans remain unknown and we may need to gain more fundamental knowledge of the mechanisms of memory formation in order to develop these protocols (Smolen et al., 2016). Our study contributes by informing the research community that 5-minute spaces in a lecture setting seem insufficient to promote knowledge retention, and that other approaches should be investigated to develop optimal spacing formats. In future studies on spaced learning during instruction, one may specifically investigate the influence of (1) the duration of spaces (2) the number of spaces (3) in relation to the duration of the retention gap. Furthermore, one should be specific about the characteristics of the setting in which the study was performed, to determine if findings can be generalizable across educational contexts. Finally, future research may combine spaced learning with other effective learning strategies such as retrieval practice and/or test-enhanced learning to further promote knowledge retention in medical education (Roediger & Butler, 2011; Karpicke & Roediger, 2008; Ayyub & Mahboob, 2017).

**Strengths and limitations**

Our experiment was embedded in an obligatory course of the medical curriculum, so some practical limitations should be noted. For sake of time and anonymity, we did not register attendees at the time of lecture. Consequently, we were unable to question students who were absent on the retention-test about their reasons for a no-show. The higher drop-out rate in the control group thus remains unexplained. Furthermore, some sort of testing effect is inherent to our pre- post-test design. We aimed to minimise the testing effect by incorporating a two-week gap between the baseline-test and the intervention, and by including new questions in our retention-test. Specific strengths of the study design should also be delineated. First, the protocol included reliable tests in which we did not observe any floor or ceiling effects. We showed an increase in overall test scores from 10% to over ~30%. However, it is hard to contextualise this result as previous literature on retention following a lecture is heterogenous, with a high variability of the moment of delayed testing, e.g. one week (Mahler et al., 2011), 4-months (Giles et al., 1982), 5-months (Razzel & Weinman, 1977), type of testing, and restudying opportunities, e.g. summarise, note-taking or self-questioning (King, 1992). Second, the analyses included narrative comments which indicated that our new spaced lecture format was well-received. The majority of students and the lecturer noted that this format increased their attention and engagement and improved their productivity. It would be interesting to investigate whether this experience of enhanced attention could be quantified, using any approach that previously assessed mind wandering and its effect on retention in a lecture context (Farley et al., 2013; Lindquist & McLean, 2011).

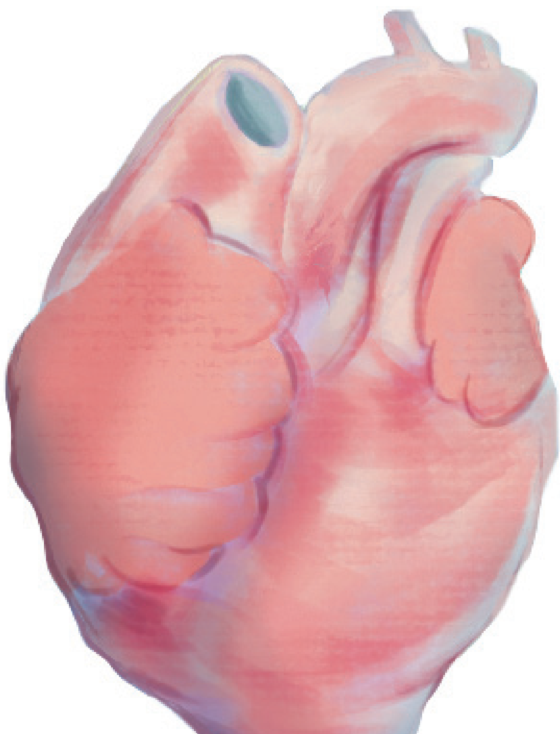
## Conclusion

Our findings showed that a spaced lecture did not enhance knowledge retention in medical students compared with a traditional, massed lecture. However, positive narrative comments indicated that this new spaced lecture format was generally well-received by students and the lecturer. Additionally, a theoretical and practical elaboration on our findings resulted in recommendations for educators and researchers on how to implement and study future spaced learning projects.

# Part 2


Concept
Learning

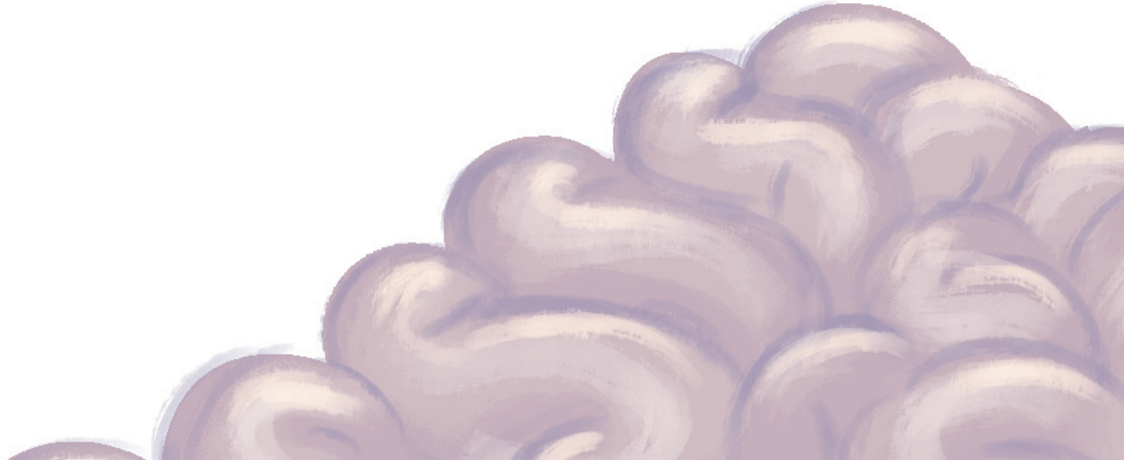
# Chapter 4



# The origins of medical students' misconceptions and misunderstandings in cardiovascular physiology

Marjolein Versteeg · Paul Steendijk

Submitted.



## Abstract

Physiology is acknowledged as a challenging subject for medical students. Many students seem to have a poor understanding of concepts underlying physiological processes. However, the specific origins of inaccurate conceptual understanding remain underinvestigated in medical education. This study explores students' fallacies in conceptual understanding regarding the interrelated concepts of pressure, flow and resistance. Qualitative conventional content analysis was carried out on medical students' ($N = 156$) written responses to a formative test. In total, seven categories emerged to describe students' fallacies; characteristics of substances, characteristics of structures, characteristics of quantities, cause-effect relationships, assumptions, systems thinking, and mathematical reasoning. This categorical distinction revealed various origins of poor conceptual understanding. These origins were linked to different educational theories on concept learning, including ontological miscategorisation and complexity-reduction. Besides these theoretical implications, our outcomes practically inform educators on potential fallacies in students' conceptual understanding that may need to be addressed. We suggest that educators may be helped by distinguishing between misconceptions and misunderstandings in their teaching. Misconceptions are robust incorrect scientific beliefs that are difficult to alleviate because the learner should undergo conceptual change, whereas misunderstandings are flaws in one's thinking that may be quickly resolved by making the learner aware of these flaws through direct instruction. Accordingly, developing conceptual change instructions and creating awareness by explicitly addressing misconceptions and misunderstandings is an important step towards developing effective instructional designs that may advance physiology education.

## Introduction

Medical education should expand students' comprehension of human body functions. This is challenging as students find it difficult to grasp an accurate understanding of the underlying physiological concepts (Michael et al., 2002; Modell, 2007). Students' conceptual understanding of medical physiology has shown to be rather limited and difficult to enhance through traditional teaching methods (Palizvan et al., 2013; Ahopelto et al., 2011). The cause may lie in the so called misconception phenomenon that refers to a learner's incorrect mental model which is incongruent with the accepted scientific model (Posner & Strike, 1982; Chi et al., 1994a).

Misconceptions in medical physiology have been demonstrated repeatedly (Michael et al., 2002; Palizvan et al., 2013; Versteeg et al., 2019). Educational theorists working on concept formation and conceptual change have provided various explanations for the existence of misconceptions and the difficulty in debunking them. For instance, students may enter their medical education with naive beliefs about concepts, based on everyday life experiences or incorrect inferences from previous education (Posner & Strike, 1982; Chi et al., 1994a; Feltovich et al., 1988). These naive beliefs appear highly resistant to change, since they may provide a system of explanation which seems relatively internally coherent (Vosniadou, 1994).

Although there is ample research on students' level of conceptual understanding in medical physiology education, the specific origins of inaccurate conceptual understanding remain underinvestigated (Fyrenius et al., 2007). In this study, our primary aim was to qualitatively explore the spectrum of students' inaccurate understanding regarding the interrelated concepts of pressure, flow and resistance. These concepts are of particular interest, since previous studies showed that medical students' ability to correctly understand these concepts and the relationships between them appeared very limited (Versteeg et al., 2019; Versteeg et al., 2019). Findings may have implications for teaching and learning of these concepts specifically. Moreover, we expect our results to enrich theoretical frameworks on concept learning in general so they may also be transferred to other contexts.

## Methods

### Context of the study

The study was carried out in the medical school at Leiden University Medical Center. One-hundred fifty-six first-year medical students ($N$ = 156) participated in this study, which was performed during a small group teaching seminar on cardiovascular physiology. This seminar was part of a compulsory, 8-weeks course on integrative cardiovascular, respiratory and kidney physiology at the beginning of the second semester. In the first semester, students had completed courses on cellular and molecular human biology. During the physiology course, students met in seminar groups (12-17 students per group) once a week for 2-4 hours. The total amount of scheduled time for the course is, on average, 20 hours/week. Besides the seminars, scheduled educational activities are lectures, practical sessions, computer sessions, and interactive sessions with experts. Summative assessments are performed half-way and at the end of the course by exams with multiple choice questions. Additionally, students earn credits by writing an essay. The seminars focus on solving clinically-based scenarios in small-group sessions led by an expert. This study was performed during the first seminar of the course. As preparation for the seminar the students were instructed to study a textbook section (Boron & Boulpaep, Medical Physiology, 3rd ed, p.412-420) on the basics of hemodynamics. Briefly, this sections handles the definitions of flow, (driving) pressure and resistance, and how these are related via Ohm's law of hydrodynamics. It explains how vascular resistance is related to the length and diameter of a vessel and the viscosity of the fluid via the Poiseuille equation, and under which conditions this equation is valid (or at least a good approximation). Furthermore, it is explained how overall resistance can be calculated in case of multiple resistances arranged in series and/or in parallel. These topics were also introduced and explained in a prior plenary lecture.

**Study design**

During the seminar, students were given a formative test consisting of eight open-ended questions. These were two-tier questions; the first tier was a multiple choice question, e.g. Q: "what happens to the bloodflow?" A: "Increases/decreases/stays the same", and the second tier was an open-ended question asking to explain one's answer to the first tier. The eight questions were divided in two sets of four, for which students received the correct explanations of the first set before answering the second set. All questions were related to the concepts of flow, pressure and resistance of blood in the cardiovascular system. Students had four minutes per question to provide an answer to the multiple choice question plus a written explanation All questions and correct explanations were designed by a physiology teacher (PS) with longstanding experience in cardiovascular research and teaching, and designing and reviewing exam questions.

### Data analysis

Using a qualitative approach, the written explanations were first independently coded by two researchers (MV and PS). A conventional content analysis was performed, which means that the codes were derived directly from the text data (Hsieh & Shannon (2005). The team members read the answers and performed open coding in isolation. The team met on a regular basis after coding each question to review the code structure and to resolve any discrepancies between codes. The coding process was iterative as codes were adjusted until full agreement was reached. For each of the four initial questions, a separate codebook was established. Subsequently, these codebooks were applied to the second set of questions and adjusted when needed. In the final stage, all codebooks were merged to identify overarching themes in the data. Educational theories on concept learning were informative during this final stage and helped to put these overarching themes into broader perspective (Chi et al., 1994a; Feltovich et al., 1988).

### Ethical considerations

Students participated in this study on a voluntary basis and they could withdraw their consent at any time. Students received no additional credit for participating in the research. They were informed that all data would be anonymised and test performance had no effect on their course grade. The study protocol was approved by the LUMC Educational Research Review Board (ERRB).

## Results

From the analysis seven categories emerged, labeled as: characteristics of substances, characteristics of structures, characteristics of quantities, cause-effect relationships, assumptions, systems thinking, and mathematical reasoning. The labeled categories are summarised in Table 1 with a description and clarified by examples. In the text, quotes from students' (S1-S156) answers to the various questions (Q1-Q8) can be found.

### Characteristics of substances

Students described a substance using characteristics that physically do not belong to that specific substance. For instance, students stated that blood, i.e. a substance: liquid, is compressed when a vessel narrows. However, in fact, liquids are incompressible unlike gases. Some argued that the blood is compressed either within or before the narrowed section, resulting in locally increased blood pressure. This fallacy was often found when the question involved vasoconstriction.

> "Blood is compressed due to the vessel narrowing, resulting in increased pressure." (S1, Q4)

**Characteristics of structures**

Students incorrectly related specific characteristics to physical structures, e.g. series and parallel vascular networks, leading to incorrect implications or effects. For instance, students stated that in case of a parallel vessel network, i.e. a structure, blood will always be equally distributed between the vessels, implying that this distribution of flow occurs independent of the vessel's characteristics. Also, some students wrote down that in a parallel vascular system, all blood will flow to the one vessel with the lowest resistance. However, in fact, blood is distributed according to the characteristics of all vessels in the system. This fallacy was often found in questions involving a parallel system of either two or three vessels of which one was partially or fully restricted.

> "By restricting one of three parallel vessels, more blood has to pass the two remaining vessels. One of the remaining vessels has a higher resistance due to its smaller diameter. Blood will choose the easy way and will flow through the other vessel…" (S56, Q2)

**Characteristics of quantities**

Students used an incorrect definition of a quantity or attributed the wrong characteristics to a quantity. They often seemed to either interchange or equate quantities that are fundamentally different and thereby assign the wrong characteristics to a specific quantity. For instance, students stated that a blood vessel has a maximum capacity [a characteristic] beyond which flow [a quantity] cannot increase because it does not "fit". These students appear to have interchanged or equated the quantities flow and volume. In fact, even in a completely stiff tube, in which a volume change would be impossible, flow will always increase when perfusion pressure increases (although the rate of increase declines when flow becomes turbulent).

> "Flow increases, but the capacity of the blood vessel is not enough to double the flow." (S68, Q3)

**Cause-effect relationships**

Students made incorrect or unjustified inferences. For incorrect inferences, cause and effect were sometimes inversed, e.g. a change in flow causes a change in resistance, instead of the other way around. For unjustified inferences, students tended to describe cause-effect relationships between two quantities, e.g. resistance and flow, while ignoring the effect of a third quantity, e.g. pressure.

> "Flow in vessel #1 has halved, causing the resistance to double." (S102, Q7)

**Assumptions**

Students made incorrect or unfounded assumptions that led to an incorrect conclusion. This usually entailed the incorrect assumption that either the flow, the pressure or the resistance remained constant and thereby ignoring the effect of its actual change, leading to an incorrect conclusion, e.g. many students stated that when one of the vessels in a parallel structure is blocked all flow previously flowing through that vessel is now rerouted through the other parallel vessels. This indicates that the students assumed that the total flow remained constant. This is generally not correct because the vessel blockage implies that total resistance increased and thus, unless perfusion pressure increased, total flow actually decreased.

> "The ΔP stays the same and the R stays the same too, so the flow does not change." (S57, Q2)

**Systems thinking**

Students mentioned that local changes only have local effects. They did not consider or recognise the context of the system and only reasoned by considering an isolated part of the system. For instance, some students thought that local alterations in the system, such as a vessel narrowing, would only have an effect on local flow or local pressure and many students argued that upstream pressures are not affected by more down-stream alterations in vessel diameter.

> "The vasoconstriction happens after the place were pressure is measured. So, at the place of the pressure sensor the resistance and flow remain unchanged, so ΔP should not have changed either." (S23, Q4)

**Mathematical reasoning**

Students showed difficulty in interpreting and applying mathematical formulas or mathematical rules. For instance, some students incorrectly interchanged

Table 1 | Categories of students' fallacies in conceptual understanding.

| Category | Description | Example |
|---|---|---|
| Characteristics of substances | Incorrect assignment of a certain characteristic to a substance. | Blood can be compressed to a smaller volume. |
| Characteristics of structures | Incorrect implications or effects of a specific physical structure. | Blood will always be equally distributed between parallel vessels<br><br>An additional resistance always causes increased total resistance. |
| Characteristics of quantities | Incorrect interpretation or definition of a quantity, or interchanging of quantities. | A vessel has a maximum capacity beyond which flow cannot increase.<br><br>If there is still space for more flow, pressure will not increase.<br><br>If a vessel is blocked its resistance disappears. |
| Cause-effect relationships | Incorrect or incomplete inferences. | Increased flow causes reduced resistance.<br><br>If resistance doubles, then pressure drop across the resistance also doubles. |
| Assumptions | Incorrect or unfounded assumptions. | Flow always remains unchanged, e.g. when a vessel narrowing occurs. |
| Systems thinking | Incorrect notion that local changes only have local effects and are independent of the system as a whole. | Pressure changes only in the area of a vessel narrowing |
| Mathematical reasoning | Incorrect application or interpretation of formulas or mathematical rules. | If resistance decreases and pressure increases then flow remains constant, i.e. the net effect is zero. |

percentages, fractions, differences and ratios. Furthermore, some students reasoned that if two variables in an equation change in the opposite direction, these effects will cancel each other out resulting in a net effect of zero.

> "R = (P decreases) / (F increases). If the denominator increases and the numerator decreases, they will cancel each other out." (S155, Q1).

## Discussion

Many medical students seem to have a poor understanding of the concepts underlying physiological processes (Michael et al., 2002; Modell, 2007; Versteeg et al., 2019). However, the origins of inadequate conceptual understanding have received limited attention. This study qualitatively explored students' fallacies in conceptual understanding. The content was focused on the interrelated concepts of pressure, flow and vascular resistance in the cardiovascular system. A qualitative analysis of students' answers to open-ended questions on these topics revealed various inconsistencies in conceptual understanding, which were sorted in seven emerging categories. Here, we discuss our findings in the light of existing educational theories on concept learning. Moreover, we elaborate on the relevance of our qualitative approach for developing effective instructional designs in order to advance physiology education.

### Educational theories on concept learning

The fallacies in students' descriptions of characteristics of substances, structures, and quantities align with existing educational theories on concept learning namely, ontological miscategorisation, static bias, and complexity-reduction.

#### *Ontological miscategorisation*

The fallacies on the characteristics of substances and structures fit the idea of Chi and colleagues that concepts can be "ontologically" miscategorised and that the misconception can be alleviated by reassigning the concept to its right category (Chi & Roscoe, 2002). For instance, the fallacy that blood can be compressed was often used as an incorrect justification for changes in blood pressure and flow. In this case, describing blood as something that can be compressed may imply that students conceive of blood as a gas which is different from the tangible category of liquids (Vosniadou, 1994). Similar ontological miscategorisations could be identified in the fallacies that flow will only influence pressure if there is not enough space in the circulatory system, or that blood does not fit, or that resistance can disappear.

*Static bias*

Furthermore, our findings related to the fallacies on the characteristics of quantities indicated that students have difficulty to distinguish between physical quantities. This aligns with Feltovich and colleagues' idea, named the static bias (Feltovich et al., 1988). In static biases, a dynamic concept such as flow is treated as a static concept such as volume leading to incorrect inferences. Equating blood flow with blood volume was done frequently by students in our study, and is a form of inconsistent thinking that has already been described in previous literature (Feltovich et al., 1993). Feltovich and colleagues (2012) have noted similar examples related to cardiac output, in which the rate of blood flow is often treated by students as changes in blood volume. Following this line of reasoning, cardiac output would increase blood volume, and blood pressure in the veins accordingly.

*Complexity-reduction*

Our findings indicate that various categories, i.e. cause-effect relationships, assumptions, and systems thinking, comprise maladaptive ways of thinking that relate to the complexity-reduction idea of Feltovich et al. (1998). According to this idea students attempt to underdimensionalise concepts, breaking them down in separate operative variables instead of dealing with the complexity of interactions. For example, the oversimplification of cause-effect relationships was clearly evident as students tended to write about one cause leading to one effect. Also, some students described multiple one to one cause-effect relationships, but did not consider the conceptual integration of the relationships. Another example refers to the unfounded assumptions students made, which generally entailed that one of the variables did not alter. This allows one to reason with only two remaining variables, thereby reducing complexity. Complexity-reduction can also be found in systems thinking, where students would often apply a local characteristic to the complete system or structure or vice versa, e.g. the total pressure drop (ΔP) remains unchanged, therefore, the ΔP of a specific part of the system also remains unchanged. According to Feltovich et al. (1988), these reductive tendencies are readily adopted not only due to cognitive ease, but also through reinforcement of the educational system. Spiro referred to this phenomenon as an unintended conspiracy of convenience in which simplification of the material in textbooks, lecture slides, assignments, etcetera, makes teaching and learning easier for both educators and students (Spiro et al., 1987). As stated by Feltovich and Spiro, complexity-reduction can form a source of misconceptions.

*Multiplicative and proportional reasoning*

Regarding mathematical concepts, our findings showed that students have difficulty understanding the differences between fractions (or ratios) and percentages. These concepts belong to the multiplicative conceptual field (Vergnaud, 1983). There is a long history of research illustrating that learners find it hard to understand such multiplicative concepts (Behr et al., 1992). It appears difficult to get a good grasp of these concepts partially because they require reconceptualization of the notion of unit and because they call for operating with composite units instead of singleton units (Sowder et al., 1988). Misconceptions may arise when the distinction between fractions and proportional reasoning is not made explicit or when these concepts are not taught in an integrated way.

In summary, our emerged categories align with various existing theoretical frameworks on concept learning. We believe they can be transferred to other physiological contexts to inform educators and researchers about the origins of poor conceptual understanding in medical physiology education.

**Practical implications**

Exploring students' conceptual understanding has yielded outcomes that may have important practical implications for teaching medical physiology. For instance, our findings illustrate the conceptual areas that are most likely difficult to understand for students and which educators may focus on during their teaching.

Furthermore, our findings indicated that students have difficulty understanding and applying various mathematical concepts such as fractions and ratios in the medical physiology context. This implies that medical education would benefit from additional reinforcement of mathematical skills. There are only few studies addressing mathematical thinking in medical school, so further research is needed to indicate students' needs (de Leon, 2018).

Finally, we would like to suggest that educational practice could benefit from a distinction between misconceptions and misunderstandings (Trotskovsky et al., 2013). Misconceptions are very robust ideas that are incongruent with current scientific views (Posner & Strike, 1982; Chi et al., 1994a). In such cases, correction requires a fundamental conceptual change and associated conceptual change instructions (Ahopelto et al., 2011; Alparslan et al., 2003). On the other hand, a misunderstanding is considered an incorrect interpretation or understanding of something that can often be corrected readily by direct instruction. In our findings, we identified categories that may fit the group of misconceptions, namely; characteristics of substances, structures and quantities, which frequently require ontological shifts. The categories that may be considered forms of misunderstandings are cause-effect relationships, assumptions, and systems thinking. To our knowledge, such a distinction between misunderstandings and misconceptions has not been described before in physiology and may form a relevant contribution to views on concept learning in physiology education.

In general, educators should create awareness among students about these common misconceptions and misunderstandings during their teaching, thereby facilitating concept learning. We believe that these outcomes do not only have direct consequences for the specific concepts that were examined in this study, but can also be transferred to other contexts in medical physiology education. Namely, having accurate mathematical knowledge or being able to distinguish between static and dynamic concepts are general skills that will also benefit scientific reasoning in, e.g., renal and lung physiology. Future studies could further investigate how these findings can be used to optimise instructional designs in the medical curriculum.

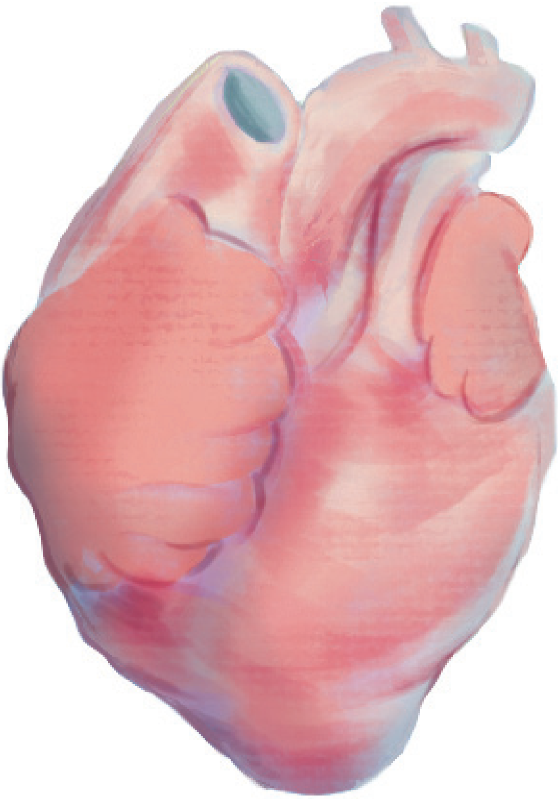
### Methodological considerations

The task was a written assignment and for practical reasons a time limit was set to answer each question. Both factors may have influenced to the completeness of responses. We cannot rule out the possibility that responses would have been more extensive and detailed if the task had been performed verbally or without time constraints. However, the written approach allowed us to perform the study with a relatively large cohort in a classroom setting that was part of the standard curriculum. We aimed for such a natural learning environment to obtain authentic responses from students that were actually enrolled in the physiology course. Furthermore, the study design included two sets of four questions each and the students were allowed to review the answers to the first set of questions before starting with the second set. This may have enhanced students' prior knowledge for this second set which may have increased the variety of responses, thereby aiding us in gaining a broader spectrum of students' understanding in this qualitative study. Lastly, a conventional content analysis, rather than a summative or directed content analysis, was chosen to gain qualitative information from student responses without imposing preconceived theoretical frameworks or categories (Hsieh & Shannon, 2005).

## Conclusion

This study reveals new insights into the understanding of physiology concepts or lack thereof by first-year medical students using a qualitative approach. A conventional content analysis on students' written responses yielded various categories that entail different origins of thinking fallacies. These categories are: characteristics of substances, characteristics of structures, characteristics of quantities, cause-effect relationships, assumptions, systems thinking, and mathematical reasoning. Acknowledgement of these different types of inadequate conceptual understanding may help educators and researchers to develop more effective instructional designs in order to advance medical physiology education.

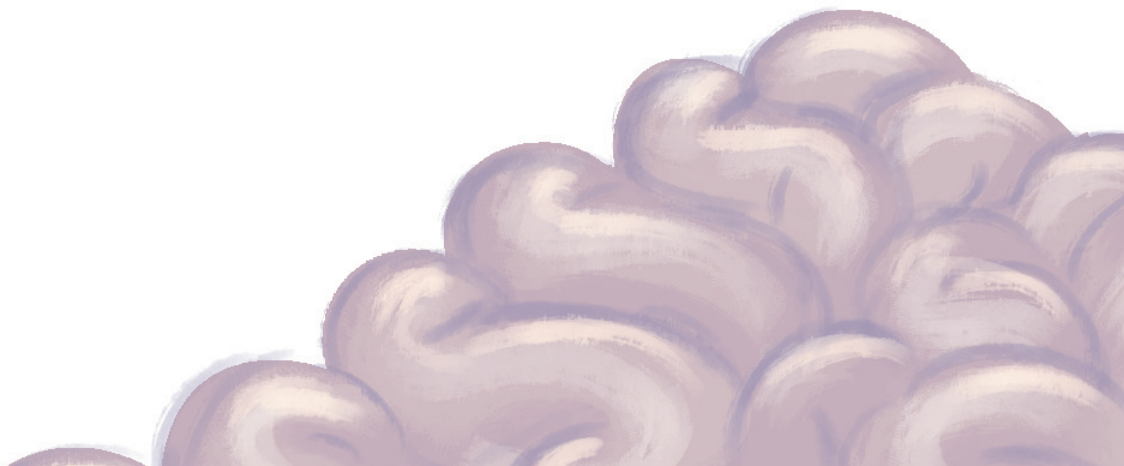
# Chapter 5



# An understanding of (mis)understanders: exploring the underlying mechanisms of concept learning using functional magnetic resonance imaging

Marjolein Versteeg · Anne Hafkemeijer · Arnout J. de Beaufort · Paul Steendijk

Submitted.



## Abstract

Obtaining adequate understanding of scientific concepts is considered challenging due to learners' erroneous misconceptions about natural phenomena. Misconceptions may coexist with scientific knowledge in the brain. Therefore, misconceptions must be cognitively inhibited in order to select the scientific knowledge. There is, however, lack of substantial neuroscientific evidence supporting this hypothesis. In this study, we sought for this evidence by investigating medical students who solved a cardiovascular conceptual problem in a magnetic resonance imaging scanner. Brain activation was compared between understanders who had the scientific knowledge, and misunderstanders who held a misconception. No significant activation was found in brain areas related to cognitive inhibition in understanders compared with misunderstanders. Therefore, we could not confirm the idea that cognitive inhibition is involved in overcoming a misconception. Instead, we found that the putamen was significantly activated in misunderstanders compared with understanders, suggesting a role for episodic memory in learners holding a misconception.

## Introduction

Learning scientific concepts can be challenging due to the prevalence of misconceptions. Misconceptions are conceptions that are inconsistent with scientific knowledge, resulting in inaccurate conceptual understanding (Chi et al., 1994a; Potvin, 2013). A learner holding a misconception is referred to as a ‘misunderstander’, whilst a leaner with the correct scientific conception is an ‘understander’. Notably, misconceptions tend to be very robust and often persist after formal education (Palizvan et al., 2013; Periago & Bohigas, 2005). In order to help educators design effective teaching and learning strategies to alleviate misconceptions, we aim to understand the underlying mechanism of concept learning.

Different theoretical frameworks attempt to delineate the learning pathway from misconception to scientific conception, also referred to as conceptual change (Duit & Treagust, 2012). A hypothesis is that initial misconceptions are never fully removed after conceptual change, but that they coexist with the scientific conception (Potvin, 2013, 2017; Shtulman & Lombrozo, 2016). The idea that a misconception can coexist with scientific knowledge implies that one should be able to select the scientific knowledge over the misconception. This process has been referred to as ‘competitive evaluation’, in which the conception with a higher cognitive utility in the context of interest is chosen (Ohlsson, 2009).

Behavioural experiments imply that competitive evaluation requires cognitive inhibition (Hung et al., 2007). This finding was based on studying reaction time as an indicator of activation of inhibitory processes. For instance, when given limited time to answer a question, students and even expert scientists tended to endorse teleological explanations of natural phenomena (Kelemen & Rosset, 2009; Kelemen et al., 2013; Shtulman & Valcarcel, 2012). The rationale behind these experiments is that one’s ability to inhibit misconceptions is impaired by processing demands. Thus, misconceptions about natural phenomena may possibly never be removed from the brain, and therefore need to be cognitively inhibited when one wants to reason in a scientifically correct way. However, there is lack of substantial evidence from cognitive neuroscience that supports the hypothesis about the role of cognitive inhibition in conceptual understanding.

Cognitive inhibition seems related to activation in specific brain areas, including the anterior cingulate cortex (ACC), anterior insula, dorsolateral prefrontal cortex (dlPFC), and parietal regions (Hung et al., 2018). However, current functional magnetic resonance imaging (fMRI) studies were unable to show significant (Family Wise Error corrected) activation in these areas in understanders who, according to the hypothesis, inhibit their misconception compared with misunderstanders who hold the misconception (Dunbar et al., 2007; Foisy et al., 2015; Fugelsang & Dunbar, 2005; Masson et al., 2014; Vaughn et al., 2020). Consequently, no study has yet confirmed the hypothesis that misconceptions are never removed from the brain and need to be inhibited when one has to provide the scientific knowledge.

Our study aims to investigate which brain areas are involved in (mis) understanding scientific concepts. Therefore, we test the hypothesis whether understanders show more activation in brain areas related to cognitive inhibition compared with misunderstanders. Our findings will provide insight in the brain areas associated with concept learning and may therefore have implications for educational practice regarding teaching scientific concepts.

## Methods

### Participants

Fifty health professions education students participated in this study. These participants were undergraduates at the Leiden University Medical Center studying medicine, biomedical sciences or clinical technology. The participants were enrolled in physiology courses during their education which were all organised by the same coordinator (PS). In these courses, all students were taught about the physiological concept of blood velocity that was underlying the fMRI task.

Students could sign up for the experiment voluntarily. In order to avoid potential priming effects, the students were not subjected to any tests or questionnaires prior to the fMRI experiment. Consequently, it was unknown if the participant was holding the misconception or not prior to the fMRI scan.

No abnormal neurological history was reported by any of the participants and they all had normal or corrected-to-normal vision. Only right-handed students were included. Written informed consent was obtained in accordance with the LUMC Institutional Ethical Review Board approval (P19.027).

### fMRI task

For the fMRI task, we used movies illustrating a common misconception in medical physiology, which entails that students believe the velocity of blood decreases when it enters a constricted section of a vessel, just like cars slow down when the road narrows (Figure 1). This thought is incorrect since blood, being a liquid, cannot be compressed and will not slow down when it enters a constricted section of a vessel. Rather, the conservation of mass law implies that a reduction in cross-sectional area goes with an increase in blood velocity, i.e. the scientific conception. All students had direct instruction about this concept during their medical training.

During the task, participants were asked to determine if movies on blood velocity were correct or incorrect by pressing one of the two buttons with the index finger or middle finger, respectively. Four different conditions were presented using four types of movies: (1) misconception movies showed the incorrect conception that blood flows slower through a narrowed part of a vessel, (2) scientific movies showed the correct scientific conception that blood flows faster through a narrowed part of a vessel, (3) positive control movies showed a vessel without narrowing in which the velocity of the blood did not change, (4) negative control movies showed a vessel without narrowing in which the speed of the blood would suddenly change. The blood velocity was illustrated by showing a single moving blood cell moving through the center of the vessel.

A participant was classified as an understander when he/she responded correctly to both the misconception and the scientific conception condition, or as a misunderstander when he/she responded incorrectly to these conditions during the fMRI task. Understanders and misunderstanders were expected to respond similarly to the positive and negative control conditions. These conditions were used to ensure that the task was understood correctly by all participants. Additionally, the negative control could potentially involve inhibition related to error-detection whereas the positive control was expected not to trigger any inhibitory mechanisms. All conditions included a vessel, but the location of the vessel was different in each movie to ensure that participants had to stay focused. Furthermore, the direction of the blood cell's movement varied from left to right and from right to left.

The stimuli were randomly presented according to an event-related design, in which the same condition never occurred twice in a row. The movies lasted 4.0 sec each. Each movie was followed by a fixation cross for 3.0-5.0 sec. In total, 82 stimuli were presented: misconception condition $n$ = 21, scientific condition $n$ = 21, negative control condition $n$ = 20, positive control condition $n$ = 20. The stimuli were divided into two runs of 41 stimuli each, resulting in a total time of 5 min 27 sec per run. The participants were given a short break in the scanner in-between the two runs.

**Procedure**

Participants were given oral instructions by the first author (MV) and watched one example movie of each condition outside the scanner. Next, participants were placed into the MRI scanner for the data acquisition. During the task, stimuli were presented using E-Prime 2.0 software (Psychology Software Tools, Inc.) via a BOLD screen LCD for fMRI (Cambridge Research Systems), which could be seen by using the mirror located in the head coil. Participants' responses (correct or incorrect) were collected with an MRI-compatible button response box. Participants were explicitly informed not to move during scanning.

After the scan session, participants were taken back to the interview room and performed a short post-test. To confirm their behavioural results in the MRI scanner, they were asked to give a written response to the question: "What happens to the velocity of the blood cell in the narrowed part of the vessel compared to its velocity before the narrowing?". Additionally, the participants provided a confidence score regarding their answer on a 5-point Likert scale (complete guess, rather unsure, doubting, almost sure, completely sure).

### Group classification

Based on their behavioural data during the fMRI task and their answer given on the written post-test, participants were classified as understanders or misunderstanders respectively. The understanders were participants who held the scientific conception that "blood flows faster through a narrowed part of a vessel". During the fMRI task, participants needed to provide 90% or more correct answers to the scientific stimuli and misconception stimuli in order to be classified as an understander. The misunderstanders were participants who held the misconception that "blood flows slower through a narrowed part of a vessel". During the fMRI task, participants needed to answer incorrectly to the scientific stimuli and misconception stimuli for more than 90% to be classified as a misunderstander. Participants who tended to be misunderstanders, but who had more than 10% correct answers for scientific stimuli and misconception stimuli were excluded from the analysis ($n$ = 2). One participant had more than 10% missing answers and was also excluded from the analysis.

### Image acquisition and preprocessing

The MRI data was acquired with a Philips 3.0 Tesla MRI scanner (Ingenia, Philips Medical Systems, Best, the Netherlands) using a standard 32-channel head coil.

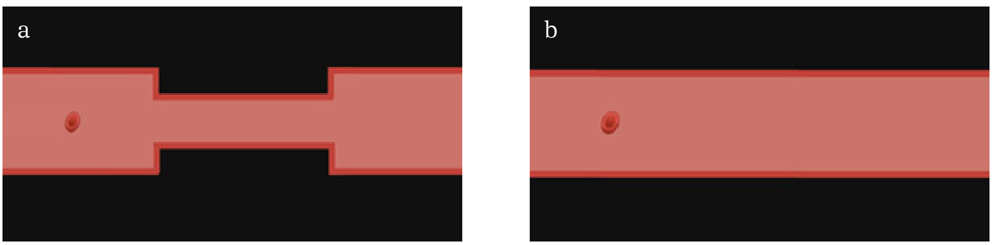



Figure 1 | Examples of stimuli presented in the MRI scanner during the blood velocity task: (a) Narrowed blood vessel used in the misconception condition and scientific conditions, (b) Control blood vessel used in the negative control condition and positive control conditions.

For registration purposes, a three- dimensional T1-weighted structural image was acquired prior to the functional images. Structural images were obtained with TR/TE = 7.9/3.5 ms, flip angle = 8˚, FOV = 250 x 195 mm, voxel size = 1.10 x 1.10 x 1.10 mm, number of slices = 155. Functional images were obtained with two runs using a gradient echo EPI sequence: TR/TE = 2200 / 30 ms, flip angle = 80˚, FOV = 220 x 220 mm, and voxel size = 2.75 x 2.75 x 2.75 mm, number of slices = 153.

Preprocessing was performed using FSL libraries (FSL 5.0.11, Oxford, United Kingdom, www.fmrib.ox.ac.uk/fsl) (Smith et al., 2004). First, non-brain tissue was removed from the structural images using a semi-automated brain extraction tool as implemented in FSL (Smith, 2002). Next, functional images were high-pass temporally filtered (100 sec), motion corrected using MCFLIRT, and spatially smoothed by using a 6-mm Gaussian kernel. Functional images were registered to the corresponding T1-weighted images using Boundary-Based Registration. T1-weighted images were registered to the 2 mm isotropic MNI152 standard space image using nonlinear registration with a warp resolution of 10 mm. All scans were submitted to a visual quality control check before and after preprocessing to ensure that no gross artefacts were present in the data.

**Data analyses**

Stimulus timing files were created for each participant based on the movie onset times and stimulus duration (= movie onset time + reaction time). To model the time course of the BOLD signal, a general linear model was used for the within- and between subject analyses (Woolrich et al., 2004). The model was convolved with a hemodynamic response function. Temporal derivatives of the convolved regressors were included in the statistical design to account for potential offsets in peak BOLD responses. In addition, six movement parameters were included in the model as regressors of no interest, to account for differences in movement, even after motion correction. Gender was included as a covariate.

FSL FMRI Expert Analysis Tool (FEAT) was used for the MRI data analyses (Smith et al., 2004). A first-level analysis within subjects was used to obtain an independent outcome for each contrast on a single subject basis. A second-level analysis was performed to concatenate the two sessions per subject (fixed effects). A third-level group analysis was performed to obtain the contrasts of interest on a group level (random effects, 5000 permutations, unpaired t-test). Group-level activation maps were created using a cluster defining threshold of $p < .001$ and a cluster extent threshold of $p < .05$ (Family Wise Error (FWE) corrected).

Participants' demographics and behavioural data were analysed using independent Students t-tests (IBM SPSS Statistics Version 23.0, IBM Corp. Armonk, NY, USA).

# Results

## Group characteristics and behavioural data

Participants were divided in two groups: understanders (*n* = 20, age: 20.4±1.9, gender: 25% male) and misunderstanders (*n* = 27, age: 21.0±1.6 ($p$ = .230), gender: 0% male ($p$ = .005)). These two groups were based on the following behavioural data: understanders answered correctly to the misconception (21 out of 21) and scientific conception (20.6 out of 21), misunderstanders answered incorrectly to the misconception (0.1 out of 21) and the scientific conception (0.3 out of 21). Both groups answered correctly to the positive control (understanders: 19.7, misunderstanders: 19.7 out of 20) and the negative control condition (understanders: 19.9, misunderstanders: 20.0 out of 20). The difference in response accuracy was significant between the groups for the misconception ($p <$ .001) and the scientific conception ($p <$ .001), but not for the control conditions. These results were fully consisted with participants' written answers on the post-test (not shown).

To verify that participants were confident in their given answer, they reported their confidence after the task on a 5-point Likert scale. There was no significant difference in the 5-point Likert confidence scores of the understanders (3.8±1.0) compared with the misunderstanders (4.0±0.6), $p$ = .320.

Table 1 presents the reaction times of understanders and misunderstanders. Overall, understanders responded slower than misunderstanders in all four conditions. Within each group, there were no significant differences in reaction times between the different conditions.

## Neuroimaging data

Figure 2 shows the brain areas that were significantly activated in understanders (Figure 2A) and misunderstanders (Figure 2B) when performing the blood velocity

Table 1 | Reaction times (sec) of understanders and misunderstanders for the four different conditions.

| Condition | Understanders (*n* = 20) reaction time *M (SD)* | Misunderstanders (*n* = 27) reaction time *M (SD)* | *p* |
|---|---|---|---|
| Misconception | 2.52 (0.14) | 2.41 (0.15) | .010 |
| Scientific | 2.51 (0.15) | 2.41 (0.16) | .040 |
| Negative control | 2.49 (0.13) | 2.40 (0.15) | .048 |
| Positive control | 2.57 (0.16) | 2.43 (0.17) | .006 |

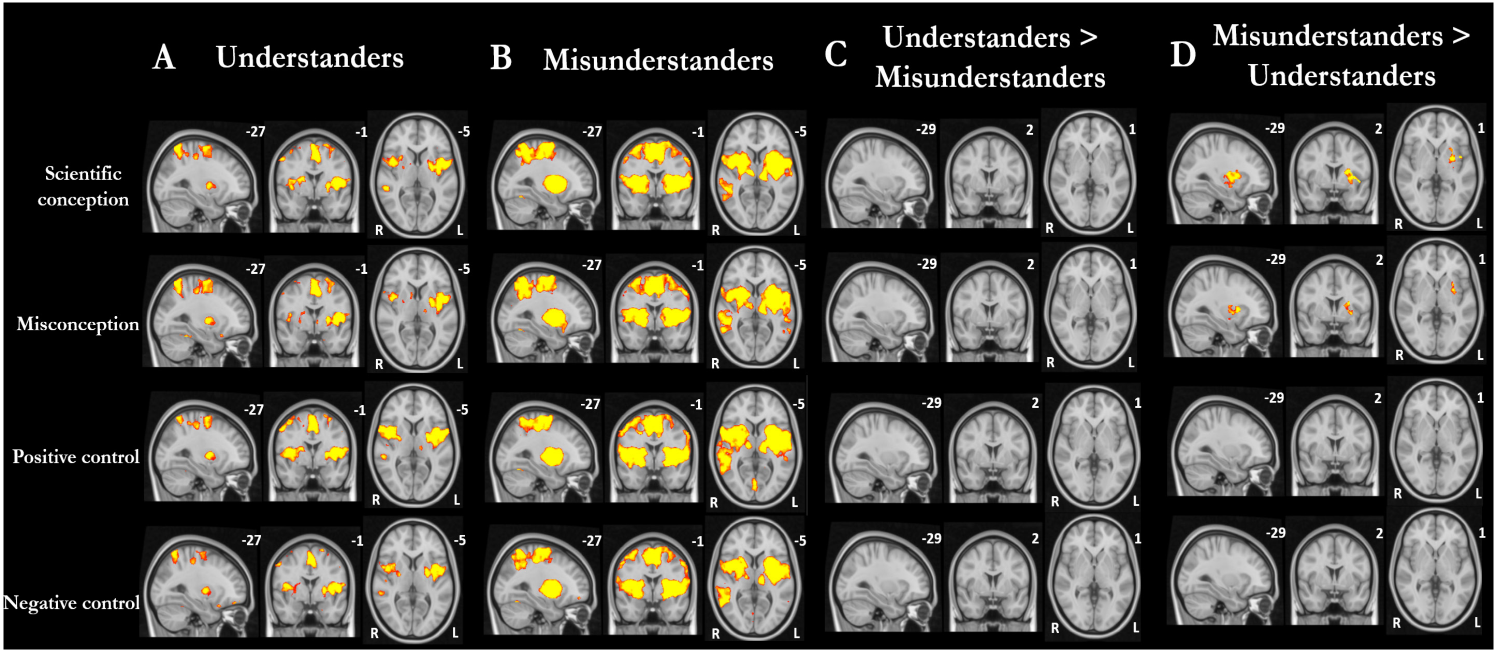



Figure 2 | Brain areas activated during the four conditions in understanders, misunderstanders, and in the comparison between both groups. A) Understanders showed significant activation in parietal and temporal areas, inferior prefrontal cortex, SMA, insula, putamen, and the upper part of the ACC in all four conditions. Only for the misconception and negative control, significant activation was shown in the anterior- and vlPFC. B) Misunderstanders showed significant activation in parietal and temporal areas, inferior prefrontal cortex, dlPFC, SMA, insula, putamen, and the upper part of the ACC in all conditions. Additionally, significant prefrontal activation in the anterior prefrontal cortex was found in the negative control only. C) In none of the conditions, understanders showed more brain activity compared with the misunderstanders. D) In both the scientific condition and misconception condition, the putamen was more activated in the misunderstanders compared with the understanders. All findings were cluster-defining thresholded $p < .001$, cluster extent thresholded $p < .05$, FWE corrected, using random effect analysis.

fMRI task. Furthermore, the comparisons between these two groups are shown (Figure 2C and 2D). Details on cluster activation can be found in Supplementary G.

*Understanders*

Within the understanders group, significant activations were found for all four conditions in superior- and inferior parietal areas (BA7, BA39, BA40), temporal areas (BA22, BA37, BA41), the inferior prefrontal cortex (BA44), the supplementary motor area (SMA, BA6), the insula (BA13) the putamen (BA49), and brain areas related to the ACC (BA24, BA32) (Figure 2A). Additionally, significant activation of the anterior prefrontal cortex (BA10) and ventrolateral prefrontal cortex (vlPFC, BA47) was shown for the misconception and negative control condition.

*Misunderstanders*

Misunderstanders showed significant activation in brain areas similar to the understanders in all conditions. These brain areas include the superior and inferior parietal areas (BA7, BA39, BA40), the inferior prefrontal cortex (BA44), temporal areas (BA22, BA37, BA41), the SMA (BA6), the insula (BA13) the putamen (BA49), and brain areas related to the ACC (BA24, BA32) (Figure 2B). Contrary to the findings in the understanders, significant activation was also found in the dlPFC (BA9) in all conditions. Additional prefrontal cortex activation (BA8, BA10) was found in the negative control condition.

*Misunderstanders compared with understanders*

Misunderstanders showed significantly more brain activation in the putamen (BA49) compared with the understanders for the scientific conception and misconception condition (Figure 2D). For the positive and negative control, there were no significant differences between the groups (Figure 2C and 2D). The understanders did not show more brain activity compared with the misunderstanders as none of the relative activations survived the $p < .05$ threshold (Figure 2C).

*Comparison of different conditions*

Figure 3 shows brain areas that were significantly activated in comparisons between different conditions. Results are displayed for understanders (Figure 3, left column) and misunderstanders (Figure 3, right column) separately. Details on

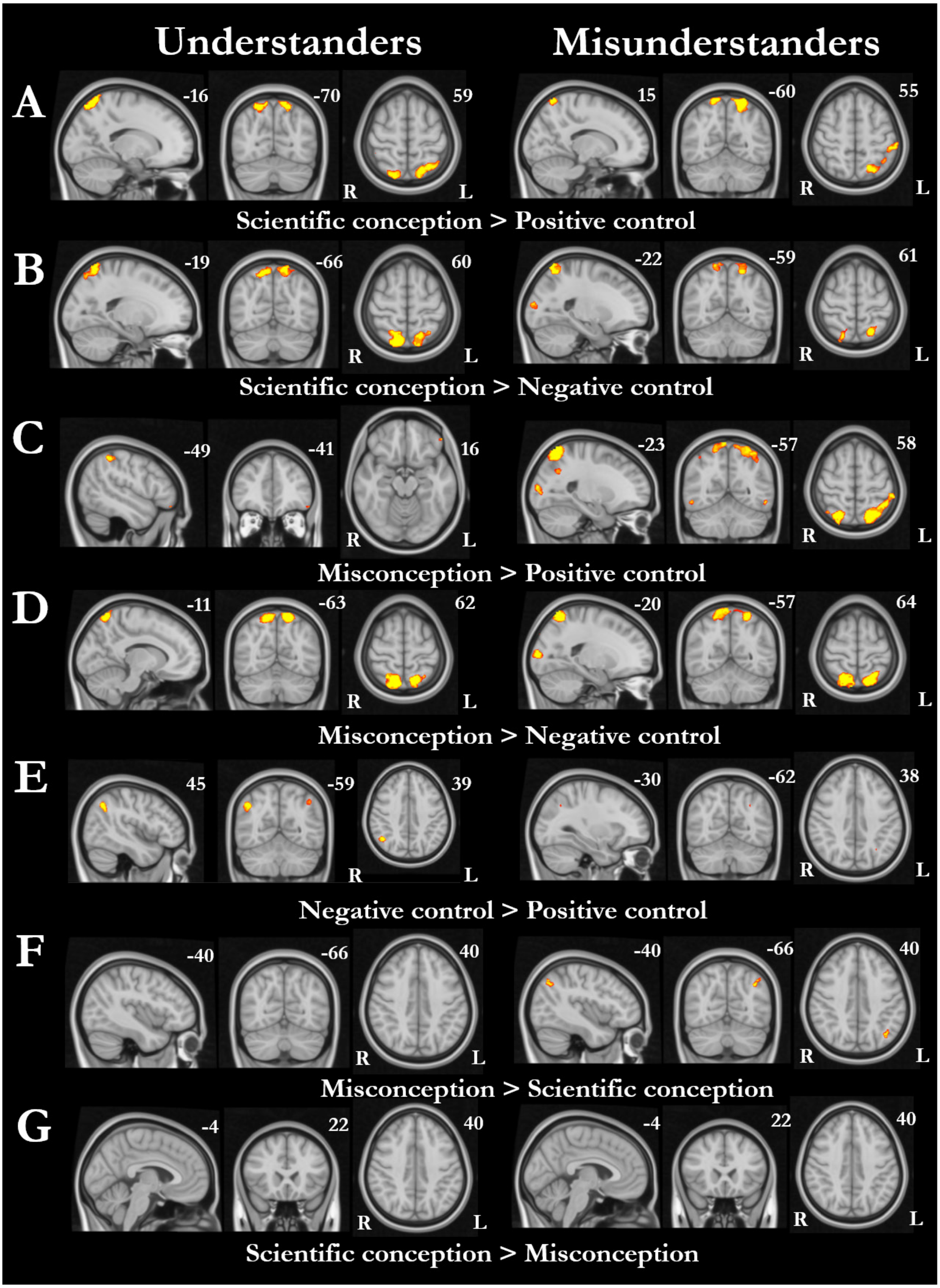



Figure 3 | Comparison of the different conditions in understanders and misunderstanders. A) In the scientific conception compared with the positive control condition, more brain activity was found in parietal regions in both groups. B) In the scientific conception compared with the negative control condition, more brain activity was found in parietal areas, visual associative areas, and the occipito-temporal cortex in both groups. C) In the misconception compared with the positive control condition, more brain activity was found in parietal areas and prefrontal areas in understanders. For misunderstanders, more brain activity was found in parietal areas, temporal areas and visual-associative areas. D) In the misconception compared with the negative control condition, more brain activity was found in parietal areas in understanders. For misunderstanders, additional greater activation was found in visual associative areas and the occipito-temporal cortex. E) In the negative control condition compared with the positive control condition, greater activation was found in parietal areas, temporal areas, and prefrontal areas in understanders. For misunderstanders, greater activation was found only in parietal areas and the occipito-temporal cortex. F) In the misconception compared with the scientific conception condition, no significant activation was found for the understanders. For misunderstanders, greater activation was found in parietal areas, visual associative areas, and temporal areas. G) In the scientific conception compared with the misconception, no significant activation was found in both groups. All findings were cluster-defining thresholded $p < .001$, cluster extent thresholded $p < .05$, FWE corrected, using random effect analysis.

cluster activation can be found in Supplementary H.

For both the scientific conception and misconception, there is more brain activity compared with the control conditions in parietal regions (BA7, BA39, BA40) in both groups (Figure 3A-D). Additionally, for misunderstanders, more brain activity was found in the visual association region (BA18) and occipito-temporal cortex (BA19) in all comparisons except for the scientific conception versus positive control (Figure 3B-D). Furthermore, misunderstanders showed more activation in the fusiform gyrus (BA37) for the misconception versus positive control (Figure 3C).

In the negative control condition compared with the positive control condition, more brain activity was found in parietal areas (BA39, BA40) in both groups (Figure 3E). Additionally, for misunderstanders, more brain activity was found for the superior parietal lobe (BA7). Understanders showed more brain activation in temporal regions (BA21, BA37, BA38).

Prefrontal brain areas displayed more brain activity in two comparisons. This finding was limited to understanders only. First, in the misconception compared with positive control, more brain activity was found in the orbitofrontal cortex (BA11) and vlPFC (BA47) (Figure 3C). Similar, more brain activity was found for the negative control compared with the positive control, with additional greater activation in the anterior prefrontal cortex (BA10) (Figure 3E).

In the scientific conception compared with the misconception, greater activation was found in visual associative areas (BA18), the angular gyrus (BA39) and temporal regions (BA21, BA37), for misunderstanders only (Figure 3F). Understanders showed no greater relative activation for this comparison.

## Discussion

The present study investigated the underlying mechanisms of concept learning by comparing brain activity in understanders and misunderstanders. The hypothesis that understanders cognitively inhibit their misconceptions could not be confirmed by our findings. Instead, we found brain activity in brain areas related to cognitive inhibition in both groups and also in control conditions. We suggest that these brain areas may play a role in memory retrieval rather than cognitive inhibition. In addition to previous studies, we report greater activation in the putamen in misunderstanders, suggesting a role of episodic memory in learners holding a misconception.

### A role for cognitive inhibition?

Our findings show more activation of brain areas that have been suggested to be associated with cognitive inhibition: the dorsal ACC (BA32), the bilateral dlPFC and vlPFC (BA9, BA44), SMA (BA6), parietal lobe regions (BA7, BA40), the insula (BA13), and the putamen (BA49) (Hung et al., 2018). However, understanders did not show more activation in brain areas related to cognitive inhibition compared with misunderstanders. Moreover, these brain areas are more activated for all conditions including the positive control condition. This makes the contribution of these brain areas to putative cognitive inhibition in this task questionable (Nenciovici et al., 2018). Consequently, we are unable to support the cognitive inhibition hypothesis that was posed by previous researchers (Foisy et al., 2015; Masson et al., 2014; Vaughn et al., 2020).

*The dorsal ACC*

Previous studies have reported on the role of cognitive inhibition in overcoming misconceptions and linked this process partly to the ACC (Masson et al., 2014; Vaughn et al., 2020). According to these studies, the ACC triggers the inhibition response when it detects a cognitive conflict between two coexisting representations. Based on our results, however, we suggest that ACC activation rather represents the monitoring of ongoing behaviour or guidance of decision-making in general (Botvinick, 2007). Namely, the ACC is more activated in all conditions in our experiment, including the positive control where no conflicting information is presented.

*The bilateral vlPFC and dlPFC*

Another brain region that has been linked to the role of cognitive inhibition is the vlPFC (BA47, BA10) (Allaire-Duquette et al., 2019; Foisy et al., 2015). Researchers found more activation of this brain area in understanders compared with misunderstanders when viewing misconception movies, and propose that the vlPFC could be directly involved in the action of inhibition (Foisy et al., 2015). In line with previous findings, we found vlPFC activation in understanders when they viewed the misconception. In our case, however, the vlPFC was also more activated in the negative control condition. Additionally, we found activation of the inferior frontal gyrus/vlPFC (BA44) in all conditions. In sum, this suggests that the role of the vlPFC in overcoming a misconception is still questionable and warrants further research. Additionally, Hung and colleagues showed that vlPFC activation is mainly associated with response inhibition instead of cognitive inhibition (Hung et al., 2018), which implies that response inhibition task (e.g. Go/No go) may be used as an additional control in experimental designs that aim to uncover underlying mechanisms of science learning. A similar line of reasoning applies to the dlPFC (BA9), which has also been shown to be involved in multiple forms of inhibition (Hung et al., 2018).

**A role for memory processes?**

We found activation of the dlPFC (BA9) in the misunderstanders across all conditions. Instead of addressing this activation to inhibitory processes, we suggest that this activation could be linked to memory processes (Brunoni & Vanderhasselt, 2014). Similarly, ventrolateral prefrontal areas (e.g. BA10) have been linked to memory retrieval by previous studies (Kim, 2010; Spaniol et al., 2009).

In addition to the prefrontal areas, increased activation in the parietal lobe areas (BA7, BA39, BA40) was found in all conditions. These regions could also reflect memory retrieval rather than cognitive inhibition (Cabeza, 2008; Cavanna & Trimble, 2006; Spaniol et al., 2009). A previous study on conceptual change by Nenciovici and colleagues also showed activation in parietal lobe areas (BA7, BA39, BA40) and related this to memory retrieval mechanisms, specifically episodic recollection (Cabeza, 2008; Nenciovici et al., 2018; Spaniol et al., 2009). Additionally, functional network studies in patients have also found the insula (BA13) to be related to episodic memory processing (Xie et al., 2012). Furthermore, the fusiform gyrus (BA37) which was active in almost all conditions has previously been associated with memory processing as well (Daniel et al., 2016) and the SMA (BA6) has been linked to working memory function (Cañas et al., 2018). Note that none of these regions was more activated in the understanders compared with the misunderstanders or vice versa, meaning that putative memory processing was induced by the task in all participants irrespective of the group they were in.

### A role for the putamen?

The only region that was found more activated in misunderstanders compared to understanders was the putamen (BA49). This region showed greater activation in the misunderstanders while scientific movies and misconception movies were presented. The putamen is involved in various types of motor behaviour. However, the brain activity was found in experimental conditions specifically, so we can rule out that this activity is related to a motor response. Potentially, the activation reflects cognitive mechanisms that are associated with the misconception movies and scientific movies. It has been shown in previous research, that the putamen is involved in learning and memory (Ell et al., 2011).

Neuroimaging studies have proposed that the putamen is associated with stimulus encoding and maintenance with its activity increasing in a cognitive load-dependent manner (Cairo et al., 2004; Chang et al., 2007). Therefore, misunderstanders may experience enhanced cognitive load when viewing scientific movies and misconception movies compared to understanders. An explanation could be that the understanders have automated their scientific response pattern to the point that the contribution of working memory has been reduced. Interestingly, however, understanders had slightly longer reaction times than misunderstanders for all conditions, which questions the idea of reduced working memory and more efficient processing among understanders.

Putamen activity has also been linked to episodic memory processes as several studies have showed that increased activity in the putamen during encoding predicts subsequent retrieval success (Ell et al., 2011; Sadeh et al., 2011; Ystad et al., 2010). Following this line of reasoning, misunderstanders may make more use of their episodic memory while viewing scientific movies and misconception movies. Perhaps this is a result of misunderstanders addressing prior naïve beliefs in order to respond to scientific and misconception stimuli. It should be noted that the putamen has been shown to be more active for semantic than perceptual episodic memory tasks (Prince et al., 2005). In sum, further research needs to be conducted in order to clarify the role of the putamen in concept learning.

### Implications for science education

It is a long way from brain imaging to the development of instructional designs for educational practice (Howard-Jones et al., 2016). Still, our findings may provide directions for teaching scientific concepts. Based on the putamen activity and the activation of other brain areas related to episodic memory during the fMRI task, we support previous recommendations that focus on the crucial role of prior knowledge in teaching scientific concepts (Hewson & Hewson, 1983). Instruction regarding conceptual understanding of scientific concepts should aim to disclose students' prior knowledge and assumptions about the concept of interest. Additionally, students' confidence in their prior knowledge may be addressed and students' awareness on the robustness of their misconceptions should be emphasised by educators. Although we did not find evidence for the role of cognitive inhibition in conceptual understanding, explicit attention to such metacognitive components of learning may also facilitate the putative control processes that are involved in concept learning (Allaire-Duquette et al., 2019; Bartley et al., 2019).

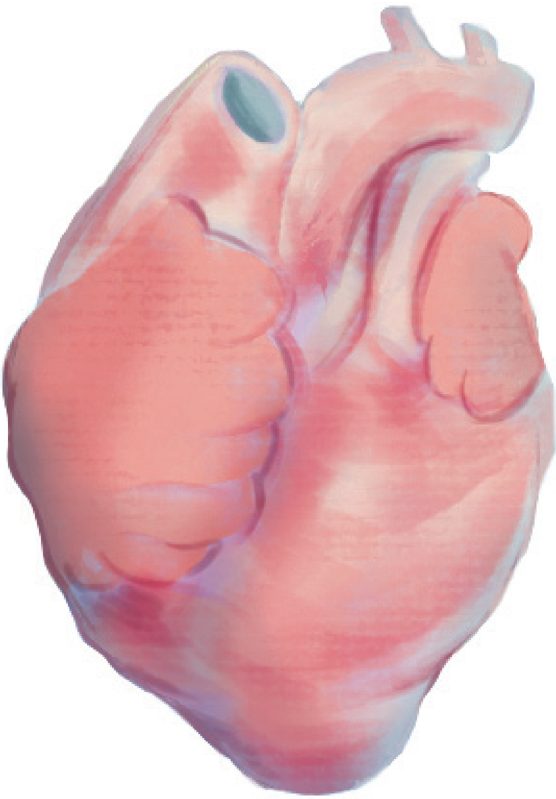
### Limitations

This study has several limitations. First, we were not sure if the understanders have had a misconception initially, e.g. prior to the physiology course. It was assumed that all participants held or have had the misconception, based on theory, but this could not be confirmed based on our study design as we did not want to induce any priming effects. The lack of insight in students' prior knowledge has also been an issue in previous research on this topic (Foisy et al., 2015; Masson et al., 2014) and should be taken into account in future studies. Second, cluster extent based thresholding is accompanied with low spatial specificity, which makes it difficult to make inferences about the statistical significance of specific locations within the cluster. Despite the low spatial specificity, this method of analysis allowed us to minimise the number of false positive findings. Lastly, only a small number of men participated in our study, which is possibly a consequence of the male:female ratio (1:3) in medical schools.

## Conclusion

A hypothesis derived from science education research states that activation of inhibitory processes is needed to overcome misconceptions that reside in the brain. The present study could not confirm this hypothesis. Instead, findings suggest a role for episodic memory in learners holding a misconception based on greater activation in the putamen. This finding supports the idea that educators should be aware of the crucial role of prior knowledge when teaching scientific concepts.

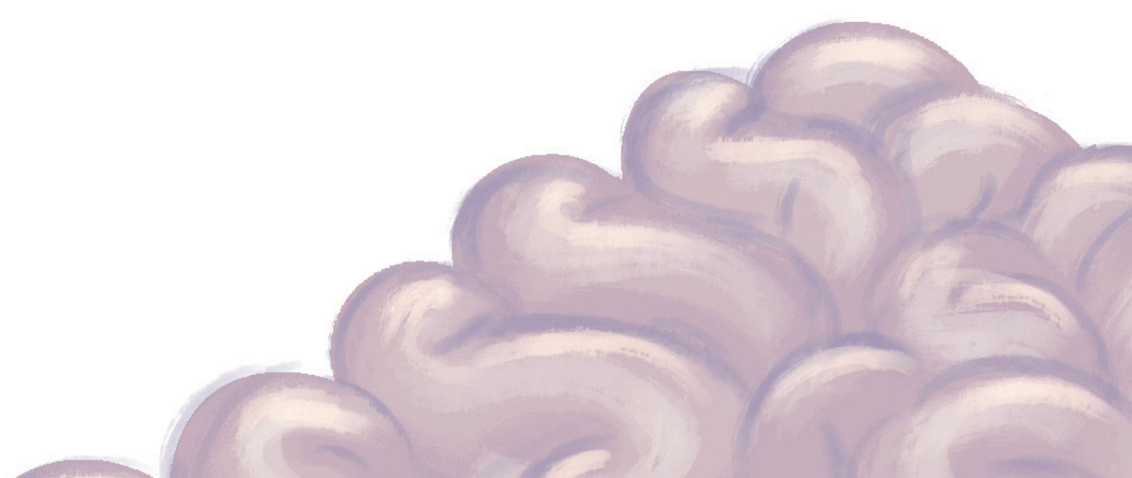
# Chapter 6



# Peer instruction improves comprehension and transfer of physiological concepts: a randomized comparison with self-explanation

Marjolein Versteeg · Floris M. van Blankenstein · Hein Putter · Paul Steendijk

Published in: Advances in Health Sciences Education (2019), 24(1), 151-165.



## Abstract

Comprehension of physiology is essential for development of clinical reasoning. However, medical students often struggle to understand physiological concepts. Interactive learning through Peer Instruction (PI) is known to stimulate students' comprehension, but its relative efficacy and working mechanisms remain to be elucidated. In this study, we investigated if and how PI could optimise comprehension of physiological concepts and transfer relative to Self-Explanation (SE) which is considered a lower-order type of overt learning. First-year medical students ($n$ = 317) were randomly assigned to either PI or SE in a prepost test design, followed by a set of near and far transfer questions. In both PI and SE groups post-test scores were significantly improved ($p < .001$) with PI outperforming SE (+35% vs. +23%, $p = .006$). Interestingly, a substantial number of students with initial incorrect answers even had enhanced scores after discussion with an incorrect peer. Both methods showed higher transfer scores than control ($p = .006$), with a tendency for higher near transfer scores for PI. These findings support PI as a valuable method to enhance comprehension of physiological concepts. Moreover, by comparing the effects of interactive PI with constructive SE we have established new insights that complement educational theories on overt learning activities.

## Introduction

Comprehension of physiological concepts is considered one of the main gateways towards efficient medical training (Finnerty et al., 2010). Conceptual insight in physiology is profoundly valuable for recall and application of clinical knowledge and diagnostic accuracy (Goldszmidt et al., 2012; Kulasegaram et al., 2013; Lisk et al., 2016, Lisk et al., 2017; Nielsen et al., 2012; Woods, 2007; Woods et al., 2007). However, acquiring comprehension of physiological concepts underlying clinical cases appears problematic for many medical students (Michael, 2007).

The presence of misconceptions is among the recurring arguments to explain the struggle that students experience in gaining conceptual understanding (Badenhorst et al., 2016; Feltovich et al., 1993; Michael, 1998). These misconceptions may originate from various sources, including teaching methodology (Badenhorst et al., 2015). An interactive instructional design named Peer Instruction (PI) has been able to improve understanding of scientific concepts (Brooks & Koretsky, 2011; Lasry et al., 2016; Miller et al., 2014; Smith et al., 2009, Smith et al., 2011; Vickrey et al., 2015; Zhang & Henderson, 2016). As developed by Mazur et al. (2001), PI contrasts with traditional lectures by including questions and subsequent peer discussions to uncover misconceptions and stimulate conceptual understanding among students (Crouch & Mazur, 2001). In multiple studies PI has shown to enhance student performance on medical physiology exercises (Cortright et al., 2005; Giuliodori et al., 2006; Rao & DiCarlo, 2000; Relling & Giuliodori, 2015), but the exact mechanisms underlying the beneficial influence of PI on comprehension remain to be clarified. In particular, studies addressing these mechanisms through direct comparison of PI with other types of overt learning strategies are scarce. This has created a gap in our knowledge about the relative efficacy of PI and in our understanding of factors that lead to the superiority of interactive learning over other overt learning methods.

Overt learning activities typically encompass student-centered instructional designs through which engagement of students in the learning process is fostered (Bonwell & Eison, 1991). The usefulness of such strategies to improve students' comprehension and tackle misconceptions has been acknowledged by the educational research community (Anderson et al., 2011; Prince, 2004). In order to clarify the relative effects of overt learning methods, Chi (2009) has developed the differentiated overt learning activities framework. According to this framework, active learning is just one form of overt learning as one can distinguish active learning from constructive learning and interactive learning. Active strategies comprise of one's engagement with the learning material without creating new knowledge, e.g. underlining sentences while reading. Constructively designed methods should promote generation of novel ideas that go beyond the presented information, e.g. Self-Explanation, whereas interactive activities involve development of each other's thoughts and consequent co-construction of knowledge, e.g. Peer Instruction. Chi has confirmed the hypothesis that interactive activities are most effective for high level cognitive processing, followed by constructive and active learning respectively (Chi & Wylie, 2014). Chi's hypothesis has been supported by educational research in the fields of physics and engineering (Lasry et al., 2016; Linton et al., 2014; Menekse et al., 2013).

Chi emphasises that the mechanisms through which interactive methods cause enhanced cognitive processing during concept learning are still unclear. Therefore, we can only hypothesise about these potential underlying processes. During constructive learning, such as Self-Explanation (SE), students are instructed to individually generate self-explanations on conceptual problems. This type of learning is suggested to stimulate one's knowledge organization and integration, which in turn is assumed to increase comprehension of the learning material since misconceptions may become more evident (Chi & Bassok, 1989). The SE effect has been demonstrated in various experimental designs, including assessments that range from memory tasks to clinical reasoning performance (Chamberland & Mamede, 2015; Dunlosky et al., 2013). Perhaps interactive learning is anticipated to be more effective than constructive learning because learners are provided with an opportunity to compare and contrast their knowledge with others, thereby establishing a more robust representation of the learned concept. Additionally, learners can build on each other's line of reasoning, a process which Chi refers to as co-construction. Note that interactive learning is thus also constructive in nature, but additionally allows individuals to expand upon their peers' thoughts and ideas. Knowing if and how an interactive strategy such as PI gains higher performance levels compared with individual constructive learning methods could help us optimise teaching and learning in medical physiology education.

The main objective (research question #1) of this study is to investigate the efficacy of interactive PI learning on comprehension of physiological concepts compared with individual constructive SE. Based on Chi's conceptual framework and a previous study from Lasry and colleagues (2016) who show superiority of PI over individual reflection in physics education, we expect that PI will outperform SE. A second objective (research question #2) is to explore the specific influence of dyad composition in the PI groups to get a better grip on the working mechanisms of interactive learning. We hypothesise that correct individuals will generally have a positive influence on their peer. Additionally, dyads that consist of students who both have incorrect answers may still increase their comphrension due to the opportunity of comparing and contrasting their thoughts and ideas, leading to potential co-construction of knowledge. Such co-construction may also lead to enhanced transfer of knowledge. Transfer has shown to be a substantial challenge for students (Kulasegaram et al., 2012, Kulasegaram et al., 2015), but the ability to apply knowledge to novel contexts is essential for developing useful clinical reasoning skills (Woods, 2007). However, the effect of PI and SE on transfer performance in medical physiology has not been addressed in previous research. Our third objective (research question #3) is therefore to study the effect of PI and SE on transfer performance as this provides insight in the extent to which constructive and interactive methods can promote learning. Based on the idea of co-construction, we hypothesise that if these learning strategies would promote comprehension, then transfer of knowledge may enhance accordingly. In all, findings on the relative efficacy of PI, its working mechanisms, and its influence on transfer will add valuable information to the active-constructive-interactive framework. Moreover, outcomes may enable researchers and educators to optimise instructional designs for concept learning in medical education.

## Methods

### Participants and setting

This study was conducted in a cohort of first-year medical students ($N = 321$) at the Leiden University Medical Center (LUMC). The study protocol was implemented in a compulsory, 8-weeks course on integrative cardiovascular, respiratory and kidney physiology. The protocol was executed in two different mandatory supervised seminars (24 groups, 12–14 students/group). Allocation to the groups was arbitrarily except for the aim to have a similar female/male ratio in all groups.

Although the seminars were mandatory, enrolment in the study was on a voluntary basis as students autonomously decided if their answers could be used for research purposes by signing the informed consent form. Students were informed that data analyses would be performed anonymously and that they could withdraw their consent at any given time. Moreover, they were ensured that the results would not affect their course grades. Students did not receive any additional credit for their participation. The study protocol was reviewed and approved by the LUMC institutional scientific committee on educational research.

### Study design

The study protocol was conducted in two different seminars spaced two weeks apart, the study flow chart is shown in Figure 1. Each seminar group was allocated to a specific intervention. In the PI groups, students were coupled with a peer to form dyads. Students were instructed to discuss their answers on the specified concept with their peer. The formation of dyads was randomised using a pseudo-random number generator and investigators were blinded to the allocation. In the SE condition, students were instructed to reflect critically on their answers in silence using the summary sheet to generate explanations and optionally write down their thoughts.

At the start of the seminar, students received a sheet summarising two physiological concepts, i.e. resistance (concept A) and compliance (concept B), that were discussed in previous plenary lectures (Supplementary H). After reviewing the information, students individually answered a set of four multiple-choice (MC) recall questions, i.e. recall test, consisting of two concept A and two concept B questions. Performance on these questions was used to compute an individual baseline score. Then, students had to individually answer four MC questions, two related to concept A and two to concept B, i.e. pre-test. After the pre-test, they were assigned to an intervention (PI or SE) which focused on only one of the two concepts followed by the same MC questions that students had to re-answer individually, i.e. post-test. By applying the intervention to only one of two concepts, an internal control for each participant was established. Subsequently, two near transfer and two far transfer questions on both concepts were performed by all students. In this study design, transfer questions are classified as congruent or incongruent to

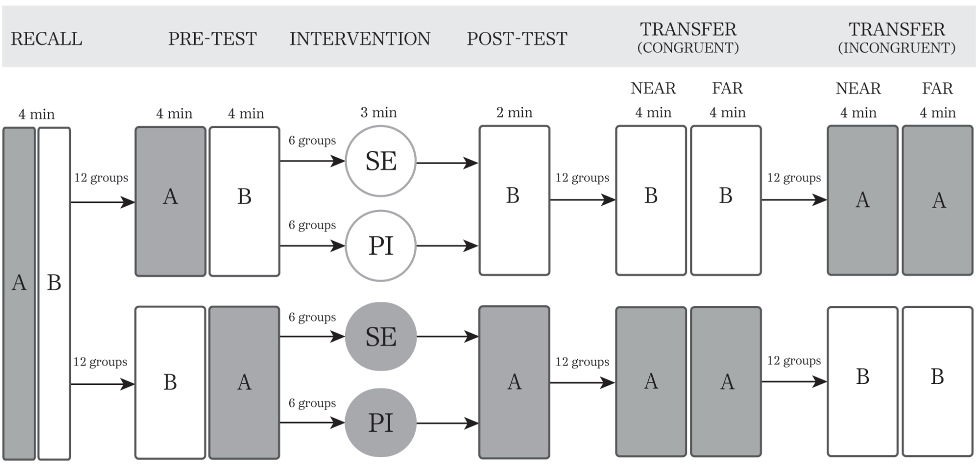



Figure 1 | Flowchart of interventions and general study design. This schematic depicts all four protocol trajectories, including six seminar groups per experimental condition ($SE_B$, $PI_B$, $SE_A$, $PI_A$). All groups received the same recall test consisting of two concept A (grey) and two concept B (white) questions. Subsequent pre-post and transfer exercises consisted of two consecutive questions each. The order of exercises in the pre-post-test design and transfer phase depends on the applied intervention. Time limits for each exercise are indicated. The same schematic was also followed in the second seminar with different exercises.

the intervention depending on one's learning intervention trajectory during the experiment. For example, a near transfer question on concept A was classified as congruent for all individuals that received PI or SE on concept A and incongruent for those receiving PI or SE on concept B. Finally, the seminar continued with a class-wide, teacher-supervised discussion and explanation of all 16 questions.

### Materials

Questions were framed in accordance with the revised version of Bloom's taxonomy of the cognitive domain; recall and comprehension questions for baseline testing, and application questions for comprehension and transfer testing (Anderson et al., 2001). Transfer questions involved similar contexts, i.e. near transfer, and novel contexts that had not been addressed during the lecture or seminar, i.e. far transfer (Barnett & Ceci, 2002). The information sheet and questions used in the protocol were designed by an expert physiologist (PS). All questions used in this study were derived from a database containing previous exam questions.

### Outcome measures and statistics

IBM SPSS Statistics Version 23.0 (IBM Corp., Armonk, New York, USA) and GraphPad Prism Version 7.02 (GraphPad Software, La Jolla, California, USA) were used for all data analyses and visualizations. Descriptive statistics are provided as means and standard errors of the mean, unless otherwise mentioned.

*Baseline knowledge*

A recall test was used to compute baseline knowledge. Comparison of the average recall scores, with a maximum score of 4 points, in the total of 24 groups would reveal possible differences in baseline knowledge prior to intervention. Similar baseline knowledge between groups ensures that intervention effects on comprehension and transfer are not due to differences in baseline performance prior to intervention. A one-way ANOVA was conducted including $SE_A$, $SE_B$, $PI_A$, and $PI_B$ as experimental groups to test baseline differences.

*Comprehension*

A comparison of post-test scores between PI and SE groups provides an indication of the relative effect of each intervention on students' comprehension. We applied an ANCOVA to determine the effect of PI and SE intervention on the post-scores, correcting for students' pre-test scores.

*Peer influence*

The specific influences of peers on comprehension in the PI group were assessed by two analyses. Firstly, we examined the probability of staying or turning correct after discussion with either an incorrect or correct peer. Outcomes were compared to SE to measure the relative effect. Secondly, we examined the change in different dyad compositions, i.e. both correct (C–C), one correct (I–C), both same answer incorrect (I–I) or both different answer incorrect (I–I*). Dyad compositions before and after PI intervention were analysed to determine the effect of peer–peer interaction on post-test performance. Computing the prevalence of these dyad compositions before and after PI intervention indicated the influence of an incorrect or correct peer on post-test performance of a participant. A Chi square test of independence was performed to determine differences in dyad composition before and after PI intervention. Continuity correction was applied for 2 x 2 contingency tables.

*Transfer*

Transfer was assessed by measuring students' performance on near and far transfer questions. All students received questions on two concepts, i.e. concept A and B, but the intervention (PI or SE) was applied to only one of these concepts. The comparative effect of PI and SE on transfer was assessed by measuring the difference in performance scores between both interventions on congruent questions. Additionally, the absolute effect of each intervention was computed by comparing students' performance on congruent and incongruent questions. A generalised estimating equations (GEE) approach was used to analyze performance on transfer exercises corrected for multiple comparisons and repeated measures. In this model, scores on near and far transfer questions were predicted based on the intervention (SE-PI), type of transfer (near-far) and protocol trajectory (congruent-incongruent).

## Results

A total of 321 first-year medical students from Leiden University Medical Center were enrolled in the medical physiology course, of whom 317 students consented to participate in the study (98.7%).

### Comprehension

Results on the recall test prior to intervention showed similar performance outcomes between the four experimental groups ($SE_A$ ($n$ = 83), $SE_B$ ($n$ = 82), $PI_A$

($n$ = 78), $PI_B$ ($n$ = 74)) with an overall average score of 2.37 ± 0.06 out of 4 points. Analysis of variance indicated no significant between-group differences ($F(3, 313) = 0.751$, $p = .522$).

Total scores (max. 2 points) for the pre-test and post-test were computed for each student, data of the two seminars were combined to compute average scores (Figure 2). The pre-test scores were similar in the PI and SE groups (PI: 0.77 ±0.04 vs. SE: 0.70 ± 0.04, $t(620) = 1.295$, $p = .196$) and increased on the post-test (PI: 1.04 ± 0.05 vs. SE: 0.86 ± 0.04, both $p < .001$). In the PI group the average score increased by 0.27 (CI: 0.20–0.35) and in the SE group by 0.16 (CI: 0.09–0.23), indicating a greater improvement of scores in the PI group ($F(1, 619) = 7.671$, $p = .006$; $\eta_p^2 = .012$). The performance gain in PI groups was shown for all questions (Supplementary I). Additionally, a control analysis showed that the outperformance of PI versus SE resulted from both more incorrect-to-correct changes and fewer correct-to-incorrect switches by individual students (Supplementary J).

**Peer influence**

The efficacy of PI presumably depends on whether you are coupled with a peer who was initially either correct or incorrect (Figure 3). A students' initial correct answer stayed correct more often when coupled with an incorrect peer (95%)

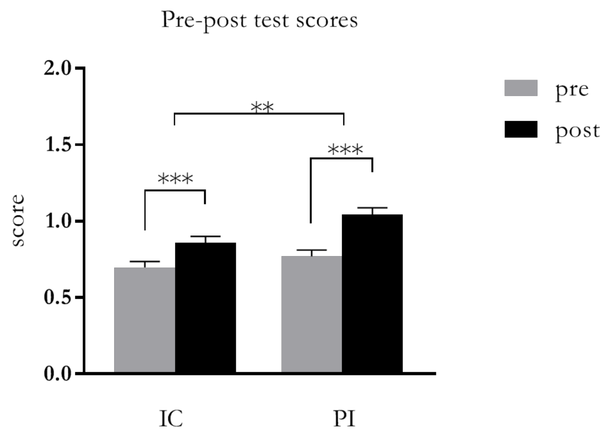



Figure 2 | Effects of PI and SE intervention on pre-post scores. An increase of 0.27 (CI: 0.20-0.35) was found for PI, compared to an increase of 0.16 (CI: 0.09-0.23) in SE. The pre-post difference in scores was thus more pronounced in the PI group ($F(1, 619) = 7.671$, **$p = .006$). Scores in the PI group (0.77±0.04 to 1.04±0.05) and SE group (0.70±0.04 to 0.86±0.04) improved significantly after intervention (both ***$p < .001$).

versus coupled with an correct peer (86%). In both cases, the chance of staying correct was higher than with SE (81%). A student with an initial incorrect answer turned to a correct answer more often when coupled with a correct peer (63%) versus when coupled with an incorrect peer (19%). With SE, 22% of initially incorrect answers were followed by a correct answer on the post-test.

Additionally, we investigated the dyad compositions to examine the influence of peer–peer interaction on post-test performance in more detail (Figure 4). There was a difference in dyad answer composition pre-intervention compared to post-intervention ($\chi^2(9) = 145.799$, $p < .001$; $\varphi = .714$). Post-intervention there were more C–C dyads (pre: 23% to post: 47%, $t(285) = 8.776$, $p < .001$). This was due to 57% of initial I-C dyads, 14% of initial I–I dyads and 18% of initial I–I* dyads that became C–C dyads after peer discussion. Also, 12% of I–I dyads and 10% of I–I* dyads became half correct, i.e. I–C, on the post-test. Furthermore, there were 6% of initial C–C dyads that changed to I-C dyads, and 12% and 8% of I-C dyads that changed to I–I or I–I* respectively.

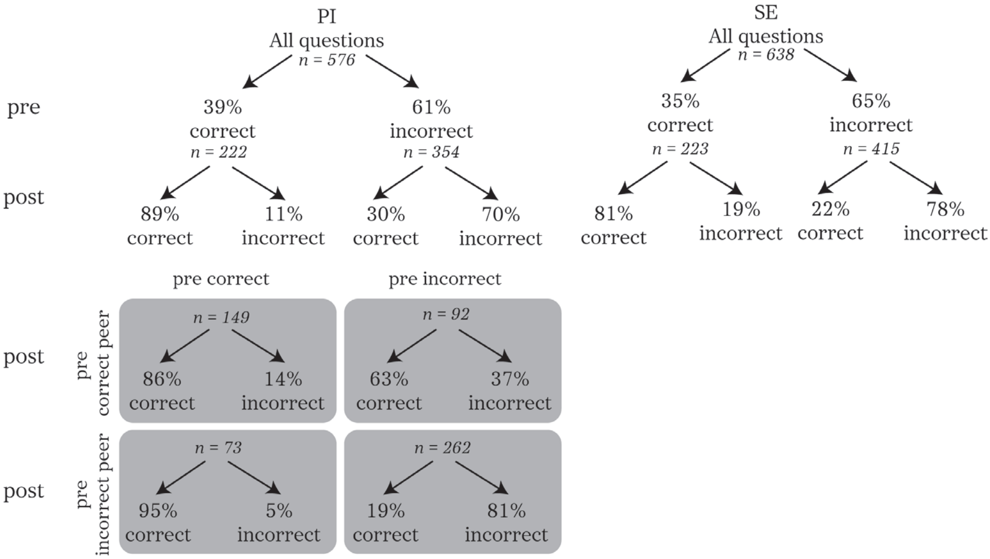



Figure 3 | Student responses on the pre-post-test. The first two rows represent the percentage correct and incorrect answers prior to intervention (pre) and after intervention (post) respectively for PI (left) and SE (right). The quadrant below visualises the answer changes for the PI intervention split into cases when the peer was correct and when the peer was incorrect on the pre-answer. Incorrect pre-answers changed to a correct post-answer most frequently after PI with a correct peer (63%) compared with PI with an incorrect peer (19%) and with SE (22%). Correct answers that were discussed with an incorrect peer stayed correct more often (95%) than after discussion with a correct peer (86%) and after SE (81%).

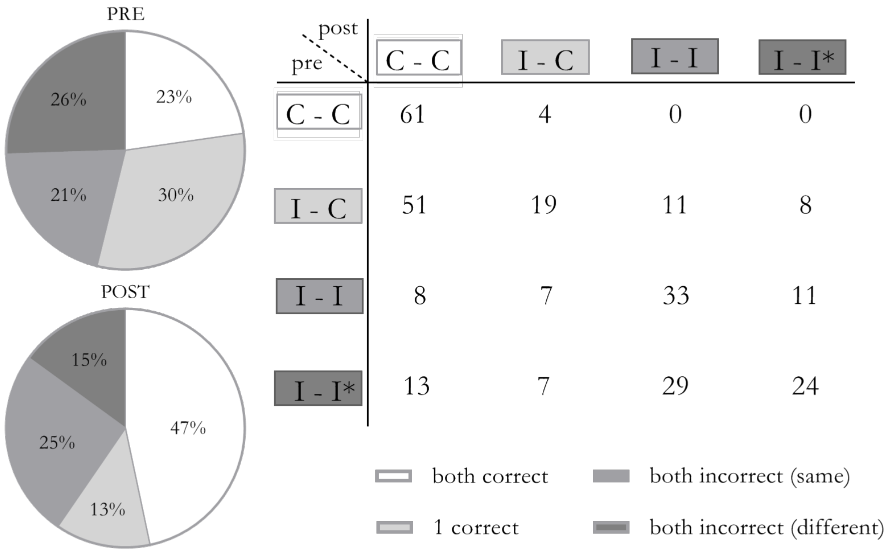



Figure 4 | Dyad compositions in the PI group before and after intervention. Pie charts show the distribution of dyad compositions pre and post intervention, with an increase of 23% to 47% correct dyads (C-C) after PI ($p < .001$). A majority of disagreeing correct (I-C) dyads pre intervention changes to correct dyads (57%, $p < .001$) and 19 I-C dyads changed to an incorrect dyad (21%, $p < .001$). Notably, 15 I-I (26%) and 20 I-I* (28%) dyads changed to either a I-C or C-C dyad after peer discussion.

**Transfer**

Students scored significantly higher on near transfer questions as opposed to far transfer questions (Table 1; $p < .001$) On both near transfer (PI: 0.95 ± 0.04; SE: 0.89 ± 0.04) and far transfer (PI: 0.80 ± 0.04; SE: 0.82 ± 0.04) tests, no significant differences in scores were found between the groups (B = 0.004, SD = 0.035, $p = .904$). Overall, students performed significantly better on congruent questions versus incongruent questions ($p = .006$). For near transfer, the difference between scores on congruent and incongruent questions, i.e. intervention versus no intervention, was 0.112 ± 0.05 for PI and 0.034 ± 0.05 for SE. For far transfer, the difference was 0.040 ± 0.05 for PI and 0.085 ± 0.04 for SE.

## Discussion

With this study we aimed to investigate the relative influence of interactive learning compared with individual constructive learning on the comprehension and transfer of physiological concepts in medical education. We found that interactive Peer Instruction (PI) and constructive Self-Explanation (SE) both

Table 1 | Predictors of transfer performances.

| Parameter | B | SD | 95% CI |
|---|---|---|---|
| Threshold | 0.353 | 0.035 | 0.000;0.284 |
| Type of transfer | -0.101*** | 0.026 | -0.151;-0.051 |
| Intervention | 0.004 | 0.035 | -0.065;0.073 |
| Protocol trajectory | 0.069** | 0.025 | 0.020;0.117 |
| Type of transfer * Intervention | -0.010 | 0.026 | -0.060;0.041 |
| Intervention * Protocol trajectory | 0.006 | 0.025 | -0.043;0.055 |
| Type of transfer * Protocol trajectory | -0.008 | 0.024 | -0.056;0.040 |

***p < .01, ***p < .0001. Dependent Variable: Answer, incorrect (0) or correct (1). Covariates: Type of transfer, near (-1) or far (1); Intervention, SE (-1) or PI (1); Protocol trajectory, incongruent (-1) or congruent (1).*

enhanced students' comprehension. The beneficial effects were more pronounced in the PI group, indicated by higher scores on the post-test and higher scores on near transfer questions. On transfer tests, interactive learning also showed enhanced performance versus the control condition, but in this respect no significant difference was found between PI and SE.

## Comprehension

This study has shown that the implementation of interactive PI in a medical physiology course significantly improved students' performance on conceptual physiology exercises. A head-to-head comparison of PI with SE demonstrated increased comprehension of students after PI (Figure 2). In addition, analysis of answer changes before and after intervention showed that PI resulted in more incorrect-to-correct switches and less correct-to-incorrect switches, yielding an overall beneficial outcome (Figure 4, Supplementary J). Results are in accordance with a previous study of Miller et al., who showed that the majority of answer changes due to PI is from incorrect-to-correct (Miller et al., 2015).

The use of active learning strategies has repeatedly illustrated to facilitate student performance and conceptual understanding in science education (Chamberland & Mamede, 2015; Freeman et al., 2014; Prince, 2004). According to the theoretical framework established by Chi et al. (2009), active learning can be subdivided into three categories, i.e. interactive, constructive and active learning (Chi, 2009a; Chi & Wylie, 2014). Chi's active learning hypothesis has already received support from experimental studies in science education (Menekse et al., 2013), and is reinforced by our findings suggesting that co-construction of knowledge through PI enhances comprehension more than individual constructive SE.

**Peer influence**

As the effect of PI on post-test performance presumably depends on the type of peer, i.e. correct or incorrect, answer changes on the pre-post-test were investigated from the peer-perspective. According to our results, a student who was initially correct benefitted more from discussion with an incorrect peer compared with a correct peer (Figure 3) and in both cases the results were better, i.e. fewer correct-to-incorrect switches, than after SE. In case a student was initially incorrect, PI with a correct peer lead to significantly more incorrect-to-correct switches than SE, and PI with an incorrect peer showed similar effects as SE. These outcomes imply that an interactive learning strategy such as PI can be preferred over constructive strategies, since even discussion with an incorrect peer yields similar or increased performance in direct comparison with SE. In agreement with Chi's hypothesis of co-construction, we suggest that an interaction between a correct and an incorrect student might yield more contrasting thoughts or ideas which drives a fruitful discussion and stimulates deeper understanding. This line of reasoning can be complemented by conceptual change research, advocating that awareness of potential misconceptions leads to increased comprehension (Piaget, 1978; Posner et al., 1982).

The influence of peer-peer interaction on comprehension was also investigated on the dyad level (Figure 4). A vast majority of dyads who were initially correct stayed correct after peer discussion and half correct dyads were able to convince their incorrect peer in most cases. Moreover, initially correct dyads never became an incorrect dyad after intervention, illustrating that students were positively influenced by their correct peer, rather than negatively influenced by their incorrect peer. Mazur supports this statement as he concluded in earlier studies that it is easier to change the mind of an individual who has an incorrect answer than it is to change the mind of an individual who initially chose the correct answer for the right reasons (Mazur, 1997).

Interestingly, a substantial number of dyads in which both peers initially had an incorrect response were still able to both provide the correct answer to the question after PI (Figure 4). We demonstrated a positive effect of peer-peer interaction on performance outcomes that was also present in dyads of which both students initially had an incorrect answer. When students had a different incorrect answer compared with their peer, their chances of providing the correct answer was slightly higher compared to students and their peers who had the same incorrect answer. These results complement previous research by Smith et al., who illustrated improved conceptual understanding after PI in a genetics course independent of students' answers in the discussion group (Smith et al., 2009). In their study, a large percentage of students reported that even incorrect students could learn from each other's reasoning as an individual could figure out the answer by talking through the question with another peer that does not know the answer either. Furthermore, a discussion with two incorrect individuals would stimulate comprehension as all options are explored and the ones that cannot be correct are eliminated. One could assume that such discussions are more explorative when students have different incorrect answers which might explain the more pronounced positive effect of peer discussion in disagreeing incorrect dyads compared with dyads consisting of two students with the same incorrect answer.

**Transfer**

We included near and far transfer questions in this study to explore the effects of overt learning interventions on transfer of conceptual knowledge. Students' performance on near and far transfer assignments did not differ significantly between PI and SE intervention (Table 1). However, in all conditions there was a better performance on transfer questions congruent to an interactive or constructive learning trajectory compared to conceptual questions on which students did not receive any intervention. This finding is in line with earlier studies reporting a positive influence of overt learning on transferring conceptual knowledge and with other studies investigating transfer in science education (Cortright et al., 2005; Kulasegaram et al., 2017; Smith et al., 2009).

Although the effects did not reach significance, the SE groups tended to perform best on the far transfer test. SE has shown to have a beneficial influence on transfer performance in previous studies (Dunlosky et al., 2013). Interestingly, the effect increases when the tasks become more complex (Chi et al., 1994b; Wong et al., 2002). Possible explanations for this observation can be derived from schema formation theory, stating that SE may enhance the connectedness of inter- and intra-schema components, thereby stimulating comprehension (Mayer, 1975).

One could argue that the higher scores on the post-test after PI versus SE result from the mere fact that two individuals might simply know more than one. However, our findings on the near transfer test suggest that, on average, students have gained better comprehension with PI since there is a tendency towards increased performance after peer discussion. Moreover, if only the 'two-heads-are-better-than-one' effect would cause the score increases on the post-questions, one would expect that the positive influence of PI would be less pronounced for difficult questions since fewer students can be matched with a correct peer.

However, this argument is not supported by our analyses as PI shows a beneficial effect for all questions (Supplementary I). Thus, other mechanisms seem to contribute to the potential of PI to enhance comprehension. In addition to the co-construction hypothesis, incorrect students may gain insight in their reasoning flaw while explaining their answer to their peer, and for correct students PI may act upon the 'teaching is learning' rationale.

**Limitations**

This study has concentrated on developing a controlled design for testing the effects of learning strategies on comprehension and transfer in a classroom setting. Due to practical considerations regarding feasibility, a limited number of questions per seminar (16 in total) were included in the protocol. In order to generalise the efficacy of active learning in medical physiology education, more concepts should be examined in a variety of near and far transfer settings. Another consequence of the chosen study design is the careful interpretation of the practical relevance of significant outcomes. An increase in test scores of 0.27 (13.5%) after PI intervention may seem minor, but is considered robust due to the large sample size and the various control conditions that were implemented in the protocol. Also, the benefits of peer discussion may become more compelling in an open question setting and by increasing the number of included questions and concepts. Lastly, although the supervising teachers were briefed and instructed to maintain exam conditions and followed a written protocol, a full guarantee that all students worked individually on the questions when they were instructed to do so cannot be given. Moreover, we cannot guarantee that each student in the PI group used the scheduled time between pre-post testing to actively discuss with their peers nor do we know if all students in the SE groups were actually generating explanations for their answers.

**Summary and future directions**

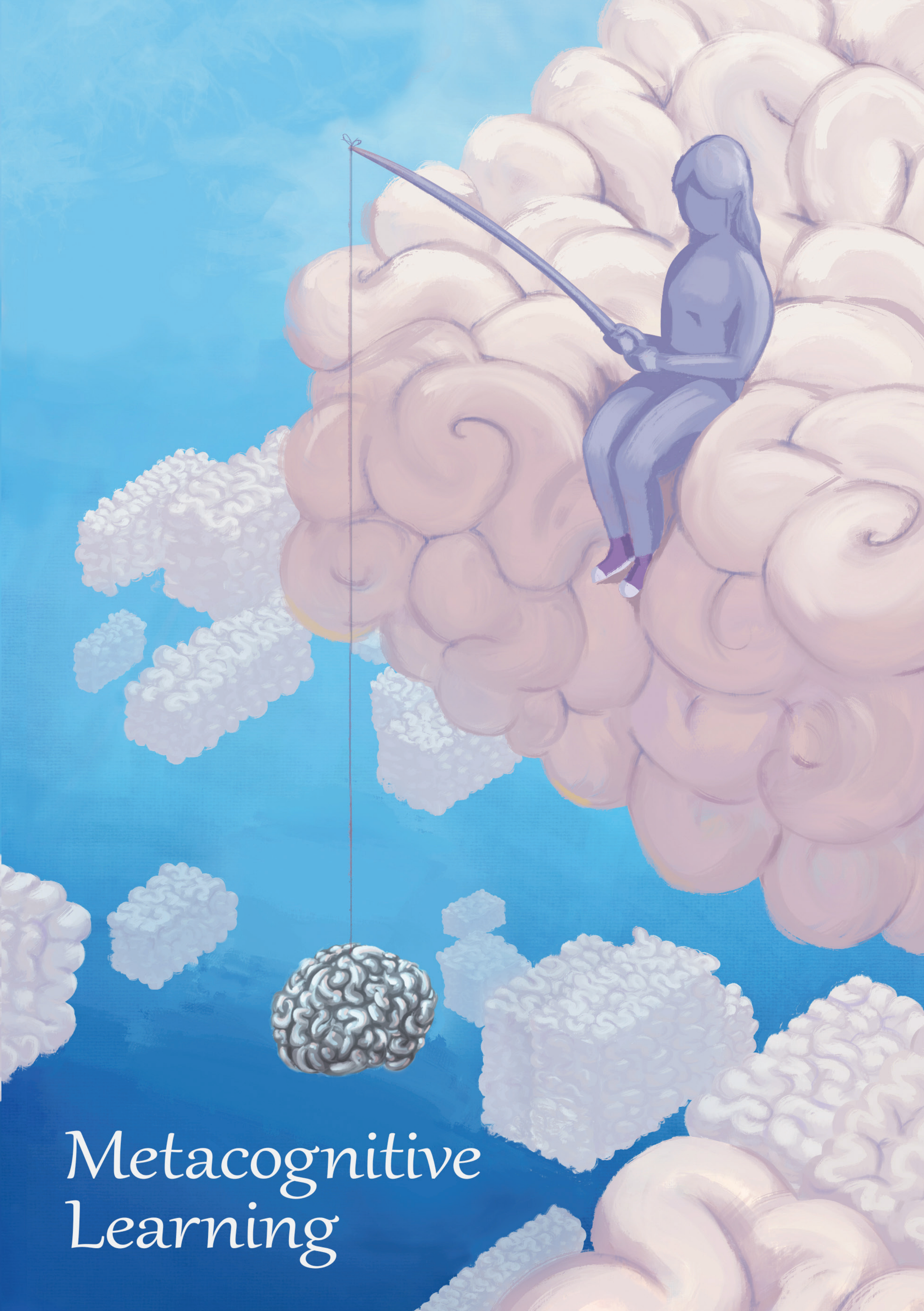
To our knowledge this study is the first to demonstrate the efficacy of PI on student performance on physiological concept questions compared with an individual constructive learning activity in medical students. Other studies have shown the additional value of PI in the classroom, but often did not address its relative effect (Cortright et al., 2005; Giuliodori et al., 2006; Rao & DiCarlo, 2000; Smith et al., 2009; Vickrey et al., 2015; Zhang & Henderson, 2016). Moreover, PI protocols are usually restricted to analysis on a group level, not confirming if the reported effects are indeed due to an increase of performance in the majority of participants (Cortright et al., 2005; Lasry et al., 2016; Rao & DiCarlo, 2000; Zhang & Henderson, 2016). By performing analyses on various levels we showed the more specific influences of PI on performance outcomes. The conclusion that even incorrect groups can achieve a correct answer after PI is in accordance with current literature (Relling & Giuliodori, 2015; Smith et al., 2009). Furthermore, research on this topic generally does not report on students' prior knowledge on the concepts of interest. By including a baseline knowledge test, we controlled for the influence of students prior knowledge on performance outcomes. Lastly, by assessment of students' performance on near and far transfer questions we investigated if the effect of overt learning interventions was transferable to novel assignments.

Future research on the efficacy of PI in medical education may include the use of open questions to gain insight in the influence of discussion on causal reasoning processes. Studies may also consider the use of between and within subject control conditions as presented in this research. Extended study protocols, including more questions, concepts and follow-up tests will provide a more accurate indication of the long-term efficacy of PI on comprehension. Also, a qualitative approach would provide additional insight in the personal thinking and reasoning processes of students and the valuable aspects of peer discussion (Brooks & Koretsky, 2011; James & Willoughby, 2011; Nielsen et al., 2016; Wood et al., 2014). From a theoretical point of view, we propose to consider the use of other theories, such as schema formation, to further uncover the working mechanisms of interactive learning strategies. Lastly, we believe that the absolute performance of students on medical physiology exercises needs further investigation. In our experiment, observations of the absolute scores indicate that a large percentage of students is incapable of achieving full conceptual understanding. Therefore, uncovering best practices or refinement of existing methods for learning and understanding medical physiological concepts should be studied in future work.

## Conclusion

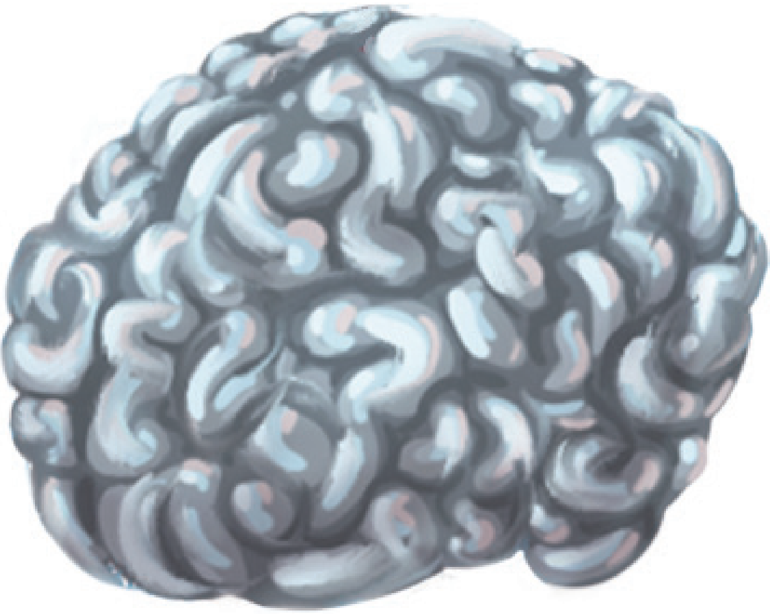
This study demonstrates the effect of active learning strategies on understanding and transfer of physiological concepts in the medical curriculum. In particular, interactive learning activities such as PI show their value for designing effective teaching methods in medical physiology education. Future research may elaborate on the working mechanisms of interactive learning activities in gaining conceptual understanding. Moreover, additional research may uncover the value of these learning strategies for transfer of conceptual knowledge to the clinical practice.

# Part 3



# Metacognitive Learning

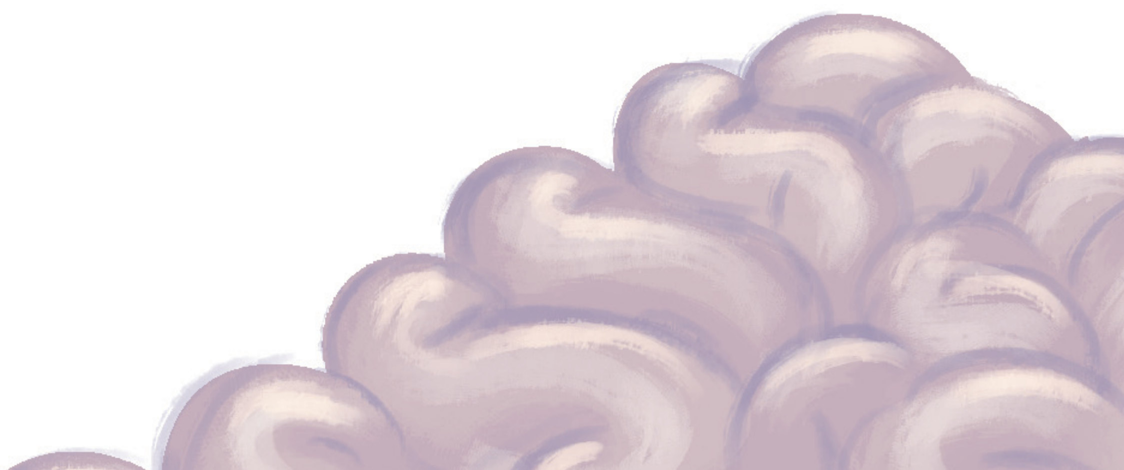
# Chapter 7



# Putting post-decision wagering to the test: a measure of self-perceived knowledge in basic sciences?

Marjolein Versteeg · Paul Steendijk

Published in: Perspectives on Medical Education (2019), 8(1), 9-16.



## Abstract

Students learn more effectively when they know what they do not know. Gaining insight into students' metacognitive awareness is needed as misalignment between actual and self-perceived knowledge impedes their learning process. The optimal method of measuring self-perceived knowledge is still under debate. In this study, we evaluate the use of psychology-derived post-decision wagering for mapping students self-perceived knowledge. Students ($N$ = 71) performed a pre-test on medical physiology, followed by a teacher-moderated discussion and a post-test with isomorph questions. Half of the students rated their self-perceived knowledge on each question using post-decision wagering, i.e. betting 1–5 points on the correctness of their answer, whereas the other half used a 5-point Likert scale to rate their confidence. Self-perceived knowledge scores were higher for post-decision wagering (pre: 3.75±0.14, post: 4.60±0.07) compared with Likert scales (pre: 3.13±0.08, post: 3.92±0.08) despite similar actual knowledge scores. Furthermore, Likert ratings showed a near-normal distribution, whereas wagers were placed preferentially using the outer ends of the scale. Correlations between mean actual and self-perceived knowledge scores were low in both groups. On average, 8.5% of responses were classified as misconceptions, defined as highly confident incorrect answers. Despite the presumed reliability of post-decision wagering, our findings suggest that we should adhere to the use of Likert scales as a balanced measure for self-perceived knowledge in medical education. Moreover, the prevalence of misconceptions did not alter after instruction, indicating a need for instructional designs that enhance students' conceptual understanding in basic sciences.

## Introduction

Learning basic sciences in medicine is recognised as an important, but challenging undertaking. It requires students to achieve understanding of complex learning material by analyzing, conceptualising, and integrating knowledge. Strikingly, students are often unaware of what they do and do not know, which can be detrimental to learning (Kruger & Dunning, 1999; Ziegler & Montplaisir, 2014). We cannot expect students to perform as effective learners if they are unaware of their own knowledge deficiencies.

Being aware of knowledge deficiencies is considered part of one's metacognition. Metacognition, put simply, refers to thinking about one's own thinking (Flavell, 1979). The ability to evaluate one's knowledge, i.e. self-perceived knowledge, after performing a task or test, i.e. actual knowledge, is a subcomponent of metacognition, which we refer to as metacognitive evaluation: knowing how well you did after finishing a task (Pintrich, 2002; Zohar & Barzilai, 2013). Metacognitive evaluation is considered a critical component for learning as it informs both students and educators on students' potential knowledge deficiencies (Cavers & Ling, 2016; Sparck et al., 2016). Moreover, some researchers have showed that including metacognitive evaluation through confidence marking during multiple-choice practice tests increases the predictive value of testing (Bokhorst, 1986; Hopkins et al., 1973; Swartz, 2006).

However, studies assessing metacognitive evaluation in medical education mainly focus on clinical knowledge, while the domain of basic sciences remains largely unexplored (Brinkman et al., 2015; Brydges & Butler, 2012; Curtis et al., 2013; Kampmeyer et al., 2015; Lai & Teng, 2011; Rangel et al., 2017; Sandars, 2009). Particularly within this latter domain, misunderstandings of physiological scientific concepts should be detected and corrected, since these concepts form a foundation for accurate clinical reasoning (Tanner & Allen, 2005; Woods, 2007). Regarding basic science knowledge, Dawson et al. (1990) illustrate that according to faculty teachers, physiological topics are experienced by medical students as being among the most difficult. This finding is supported by research that has shown a high prevalence of misunderstandings among students in medical physiology education (Badenhorst et al., 2016; Michael, 2007; Michael et al., 2002; Versteeg et al., 2019). According to conceptual change theory, these so-called misconceptions are different from a mere lack of knowledge; misconceptions are robust to change and are therefore difficult to alleviate by traditional teaching methods (Posner et al., 1982). Misconceptions are generally indicated by incorrect answers that are given by students with high confidence, indicating that students are unaware of their incorrect knowledge (Hasan et al., 1999). The role of awareness in establishing conceptual understanding illustrates that conceptual change theory and metacognitive theory are strongly intertwined in practice. Recognising and evaluating existing conceptions, and deciding whether to reconstruct or review one's understanding are all metacognitive processes (Gunstone & Mitchell, 2005). Conceptual change theory emphasises the key role of awareness in accomplishing the shift from a misconception to the scientifically correct conception (Lin & Zabrucky, 1998; Posner et al., 1982; Schraw & Dennison, 1994; Tippett, 2010). Awareness can be created by explicitly assessing students' metacognitive evaluation in the classroom, as it forces both learners and educators to think about their understanding of the subject matter.

Research on science learning in higher education nicely illustrates how metacognitive evaluation can be measured in classroom settings using a multi-tier approach (Caleon & Subramaniam, 2010a, 2010b; Sreenivalusu & Subramaniam, 2014; Yan & Subramaniam, 2018). Multi-tier assessments consist of multiple-choice exercises with a complementary Likert scale question: 'How confident are you that your answer to this question is correct?'. This is an example of measuring metacognitive evaluation; assessing students' actual knowledge through test scores and their self-perceived knowledge through Likert scales. The use of Likert scales as a self-report tool for self-perceived knowledge is very common in educational research, with variations in scales ranging from 3 to an increasing number of integers (Brandriet & Bretz, 2014; Brinkman et al., 2015; Caleon & Subramaniam, 2010a; Curtis et al., 2013; Kampmeyer et al., 2015; Lai & Teng, 2011; Rangel et al., 2017; Sreenivalusu & Subramaniam, 2014; Yan & Subramaniam, 2018). However, there is some controversy on the objectiveness of Likert scales for measuring self-perceived knowledge. As noted in a review by Koch and Preuschoff (2007): 'Studying consciousness is difficult because asking subjects to report their awareness of a stimulus perturbs that awareness'. In psychology research, an alternative method to more implicitly measure awareness of visual stimuli was introduced, called post-decision wagering (PDW) (Persaud et al., 2007). This method is suggested to determine awareness more objectively by indirectly assessing individuals' self-perceived knowledge by betting points (or money) on their decisions. If an individual chooses the correct answer points are gained, whereas with a bet on an incorrect answer points are lost. Additionally, researchers suggest that PDW enhances individuals' motivation to reveal their confidence in their answer compared with self-reports on numerical confidence scales (Dienes & Scott, 2005; Wierzchon et al., 2006). Contrastingly, various studies show a substantial influence of loss aversion in PDW resulting in a decreased accuracy towards measuring self-perceived knowledge. Therefore, participants are less prone to indicate a high confidence level through wagering

(Dienes & Seth, 2010; Sandberg et al., 2010). Moreover, wagers are also suggested to depend on wager size indicating that PDW is still a subjective tool to measure awareness (Fleming & Dolan, 2010).

The use of PDW as a tool to measure self-perceived knowledge has not yet been reported in medical education. In this study, our primary objective was to evaluate psychology-derived PDW as a measure of self-perceived knowledge in educational contexts. We compared PDW with Likert scales and hypothesised that confidence wagers rather than ratings would be more aligned with students' actual knowledge due to their supposed more objective nature. To further determine the practical usefulness of the instruments in quantitating self-perceived knowledge, we examined the distribution of responses along the scales to compare effective resolution and discriminative power. As a secondary objective, we investigated the prevalence of misconceptions among students to gain insight into their conceptual understanding of the subject matter. In the present study, we focused on basic sciences particularly given the high prevalence of misconceptions (Badenhorst et al., 2016; Dawson-Saunders et al., 1990; Michael, 2007; Michael et al., 2002; Palizvan et al., 2013; Versteeg et al., 2019). Misconceptions may not only affect students' conceptual understanding, but also have significant impact on their clinical reasoning skills (Michael, 2007; Lisk et al., 2016; Goldszmidt et al., 2012), and are therefore considered an important subject of investigation in medical education.

## Methods

### Participants

A cohort of first-year bachelor Biomedical Sciences students (*N* = 71) from Leiden University Medical Center participated in this study. In the Netherlands, similar physiology courses are part of both the biomedical and medical bachelor curriculum.

### Setting

For this study, the online platform Socrative was used to develop multiple-choice knowledge tests. The knowledge tests were implemented in a compulsory 2-hour supervised seminar on cardiovascular physiology that was taught in five small working groups (13-15 students/group) by the same teacher. Allocation of the students to the working groups was arbitrary except for the aim to have a similar female/male ratio in all groups. The topics were introduced in a preceding plenary lecture, and students were provided with a handout during the seminar which contained some basic figures and diagrams. Although the knowledge tests were a mandatory part of the seminar, students could voluntarily decide if their anonymous answers could be used for educational research purposes by giving informed consent. No incentives were offered for participation and test results did not affect the course grade. Students could withdraw their permission at any time. This experiment was part of a larger study also investigating the influence of peer discussion on confidence of which data are not reported in this paper. This study was approved by the Educational Research Review Board (ERRB) of Leiden University Medical Center: ERRB reference number: OEC/ERRB/20180612/3.

**Procedure**

The knowledge tests were taken individually under exam conditions, at the beginning (pre) and at the end (post) of the seminar. Both tests consisted of 10 multiple-choice questions and examined the same knowledge base. The post-test used slightly different phrasing of the questions which can be considered isomorphic or so-called near-transfer questions. All students were instructed to answer each question individually and to provide a confidence rating (Likert scale: 3 out of 5 groups, $n = 42$) or wager (PDW: 2 out of 5 groups, $n = 29$) immediately after each question. To compute the actual knowledge, students received one point for every correct answer of the multiple-choice tests yielding a maximum score of 10 points per test. During the body of the actual seminar the questions from the pre-test and related topics were explained and discussed.

Questions were framed according to the revised version of Bloom's taxonomy of cognitive domain and all categorised as comprehension-type questions. To ensure validity of the knowledge test, the physiology questions were designed by an expert physiologist (PS) and derived from a database of previously used exam questions. PS is an educational professor at Leiden University Medical Center and actively involved in shaping the biomedical and medical curriculum regarding medical physiology education, and familiar with the learning goals of the human biology course. Consequently, the questions were designed in such a way that they would meet the overall course objectives regarding physiological knowledge.

**Instruments**

For confidence rating in the Likert scale group (LS group), students used a 5-point Likert scale rating instrument: (1) Completely unsure (just guessing), (2) Rather unsure, (3) Neutral (50/50), (4) Rather sure, (5) Very sure (almost 100%). Students received additional instruction on the neutral item (3) as the lecturer indicated that this rating should be given when a student doubted between two remaining answer choices considered equally likely. In the PDW group, students were instructed to place their bets on the correct answer. In order to compare results with the LS group, wagers ranged from 1 to 5 points. Students could bet 1–5 points per question, which were gained if the answer was correct and lost if the answer was incorrect. Students received their total wagering scores after the seminar.

**Data analyses**

Descriptive statistics are provided as means and standard errors of the mean, unless otherwise mentioned. Actual and self-perceived knowledge scores in PDW and LS groups were compared using independent samples t-tests. Comparison of score gains on the pre-post-test between groups was performed with a repeated measures ANOVA. To test and compare alignment between actual knowledge and self-perceived knowledge scores, correlation analysis was performed using Spearman's rank order correlation coefficient. Scores were averaged per student. A Fisher's r-to-z transformation was performed to test for potential significant differences between correlations. The discriminative power of both methods was tested by performing Chi-square tests of independence. Hasan's decision matrix was used as a template to analyze the prevalence of misconceptions (Figure 1; Hasan et al., 1999). To create dichotomous outcomes, a threshold value of 3 was adopted and used as a cut-off point. Students were instructed to rate an answer with 3 when they were in doubt between two answer choices considered equally likely, meaning that they were still unsure about knowing the correct answer, i.e. having low confidence. Correct answers self-rated with a high confidence (>3) were classified as 'correct knowledge' and when rated with low confidence (≤ 3) as 'lucky guess'. Incorrect answers with a low confidence rating were considered 'lack of knowledge' and when paired with high confidence, the response indicated

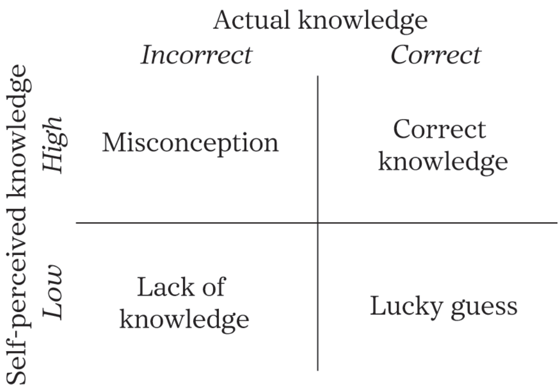



Figure 1 | A decision matrix based on students' actual and self-perceived knowledge. Adapted from Hasan et al., 1991.

the presence of a misconception. IBM SPSS Statistics Version 23.0 (IBM Corp., Armonk, New York, USA) and GraphPad Prism Version 7.02 (GraphPad Software, La Jolla, California, USA) were used for all data analyses and visualizations.

## Results

The average confidence scores reported in the PDW group on the pre-test and post-test were significantly higher compared with the scores in the LS group (both $p < .001$) (Table 1). The gain in confidence scores in the PDW group (0.86 ± 0.12) was significantly higher than in the LS group (0.79 ± 0.12) ($p < .001$, $\eta^2 = .349$). The average test scores did not differ significantly between PDW and Likert scale on the pre-test ($p = .338$) and post-test ($p = .065$). In the PDW group, students showed an average increase of 1.55 ± 0.30 points in their pre- to post-test scores, which was not significantly different from the gain of 1.32 ± 0.28 points in the LS group ($p = .094$).

### Alignment

There was a significant correlation between actual and self-perceived knowledge scores in the LS group on the pretest ($R^2 = .16$, $p < .01$) and post-test ($R^2 = 0.25$, $p = .001$) (Figure 2). In the PDW group correlations between actual knowledge and self-perceived knowledge are absent on both tests (pre: $R^2 = .12$, $p = .063$, post: $R^2 = .10$, $p = .102$). To test for potential differences between the correlations in the PDW and LS group for the pre-test and post-test respectively, correlations were transformed into z-scores using Fisher's r-to-z transformation. The differences between correlations on the pre-tests ($z = -.23$, $p = .817$) and posttests ($z = -.90$, $p = .370$) were not significant.

Table 1 | Mean scores for students' actual and self-perceived knowledge.

| Instrument | Actual knowledge (Score, max 10) | Self-perceived knowledge (Score, max 5) | No. of students |
|---|---|---|---|
| **Pre-test** | | | |
| Likert-scale (LS) | 6.38±0.27 | 3.13±0.08 | 42 |
| Post-decision wagering (PDW) | 6.78±0.27 | 3.75±0.14 | 29 |
| Post-test | | | |
| Likert-scale (LS) | 7.68±0.27 | 3.92±0.08 | 41 |
| Post-decision wagering (PDW) | 8.31±0.23 | 4.60±0.07 | 29 |

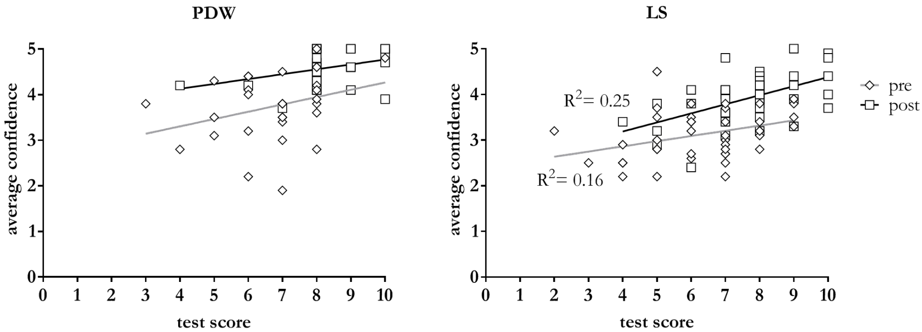



Figure 2 | Relationship between actual and self-perceived knowledge of students using PDW and LS.

## Distribution

The distribution of confidence scores (1–5) for the Likert scale and PDW instruments is shown in Figure 3. Comparing the distribution of 1–5 responses on the pre-test yielded a significant difference in distribution in self-perceived knowledge scores between PDW and Likert scale on questions with an incorrect response ($p < .007$, $\varphi = .240$). Of all questions answered incorrectly in the PDW group, 25.5% of responses were rated with level 5 confidence, compared with only 6% in the LS group. An even more pronounced difference between the Likert scale and PDW was found for correctly answered questions, yielding 43.8% level 5 confidence scores in the PDW group versus 16.3% in the LS group ($p < .001$, $\varphi = .483$). Discrepancies in the self-perceived knowledge distribution between Likert scales and PDW were also obtained for the post-test. The largest effect size was obtained for incorrect answers on the posttest ($p < .001$, $\varphi = .593$) with 53.1% level

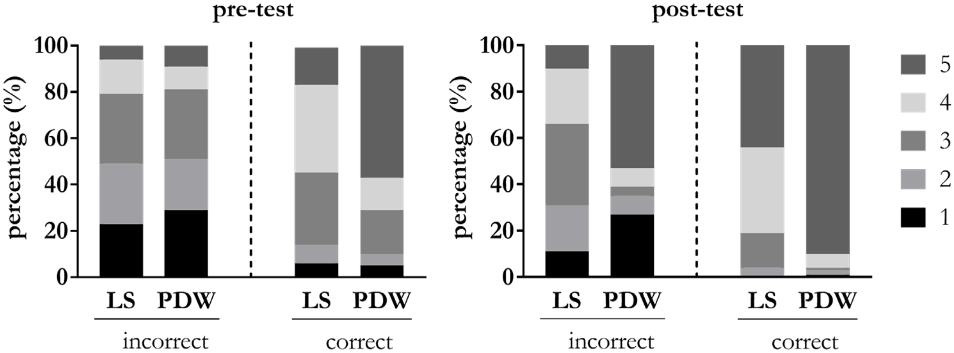



Figure 3 | Distribution of self-perceived knowledge scores.

5 responses in the PDW group and 9.7% in the LS group. For correct answers, 89.6% was associated with confidence level 5 in PDW versus 43.6% in the Likert scale ($p$ < .001, φ = .483).

**Misconceptions**

There was an increase in ‘correct knowledge’ in both groups after the plenary instruction and discussion (PDW: 50.3 to 79.7%; LS: 34.7 to 61.7%) (Figure 4). The total number of ‘lucky guesses’ was reduced (PDW: 17.2 to 3.5%; LS: 29.1 to 15.3%) as was the amount of ‘lack of knowledge’ (PDW: 24.1 to 6.6%; LS: 28.6 to 15.3%). The prevalence of misconceptions was similarly present before and after instruction (PDW: 8.3 to 10.3%; LS: 7.5 to 7.7%).

## Discussion

The aim of this study was to evaluate psychology-derived post-decision wagering (PDW) as a measure of self-perceived knowledge by comparing it with the commonly used Likert scales. Despite similar levels of actual knowledge in both groups, students who used wagers indicated more confidence in their answers compared with students who used ratings. PDW confidence scores were also less evenly distributed compared with Likert scales and did not show a normal distribution, resulting in a less proportional use of the instrument. Contrary to our a priori hypothesis, PDW thus represents a less balanced measure of self-perceived knowledge than the traditional Likert scale.

Misalignment between actual and self-perceived knowledge was present in both groups, indicating that students generally do not know what they know or do not know. Some students using PDW mentioned that they would easily go ‘all in’ and bet all their points even when they were not completely sure, and despite

**PDW**

| Self-perceived knowledge | Actual knowledge: *Incorrect* | | Actual knowledge: *Correct* | |
|---|---|---|---|---|
| *High* | pre 8.3% | post 10.3% | pre 50.3% | post 79.7% |
| *Low* | pre 24.1% | post 6.6% | pre 17.2% | post 3.5% |

**LS**

| Self-perceived knowledge | Actual knowledge: *Incorrect* | | Actual knowledge: *Correct* | |
|---|---|---|---|---|
| *High* | pre 7.5% | post 7.7% | pre 34.7% | post 61.7% |
| *Low* | pre 28.6% | post 15.3% | pre 29.1% | post 15.3% |

Figure 4 | Prevalence of misconceptions.

the rule that points were lost in case of an incorrect answer. Contrary to these findings, psychological research and behaviour economics report that individuals who wager generally show lower confidence compared with verbal confidence reports due to socalled risk aversion (Dienes & Seth, 2010; Fleming & Dolan, 2010). The apparent absence of risk aversion in our students might be due to the use of imaginary ‘points’ as incentives. Interestingly, however, studies in laboratory settings have reported similar responses when using real versus imaginary incentives (Persaud et al., 2007). We suggest that the educational environment in which students are socially engaged might influence their response to risk aversion. The effects of classroom versus laboratory settings on the use of self-perceived knowledge measures await further investigation.

Our findings are in line with previous studies reporting discrepancies between students’ actual and self-perceived knowledge consistently across disciplines (Bell & Volckmann, 2011; Cavers & Ling, 2016; Favazzo et al., 2014; Lai & Teng, 2011; Lindsey & Nagel, 2015; Potgieter et al., 2010; Versteeg et al., 2019; Ziegler & Montplaisir, 2014). Kruger and Dunning (1999) have suggested that incompetence deprives an individual of the ability to recognise its shortcomings. Based on this reasoning, our reported underestimation of performance is an unexpected finding. An explanation for this discrepancy may be that metacognitive evaluation of the specific materials in this study did not resemble the more global self-assessments mentioned in Kruger and Dunning’s studies. Such global judgements are, for example, self-competence (‘I feel I am able to …’) and self-efficacy (‘how confident are you that you can …’). These are prospective judgements and can therefore be considered part of one’s metacognitive knowledge, whereas metacognitive evaluation (as measured in the present study) is an on-line judgement which is made after one has performed a specific task (Zohar & Barzilai, 2013; Pintrich, 2002). Students might be more cautious with estimating their actual scores after they have performed a task compared with a situation in which a global prospective judgement is requested. Furthermore, gender differences may have influenced our outcomes. Research has indicated that women tend to underestimate their academic capacities, as illustrated by the study by Ravesloot on progress testing in medical education (Ravesloot et al., 2015). Most of the students (70%) in this experiment were female, which might partly explain the relatively large percentage of correct low-confidence responses.

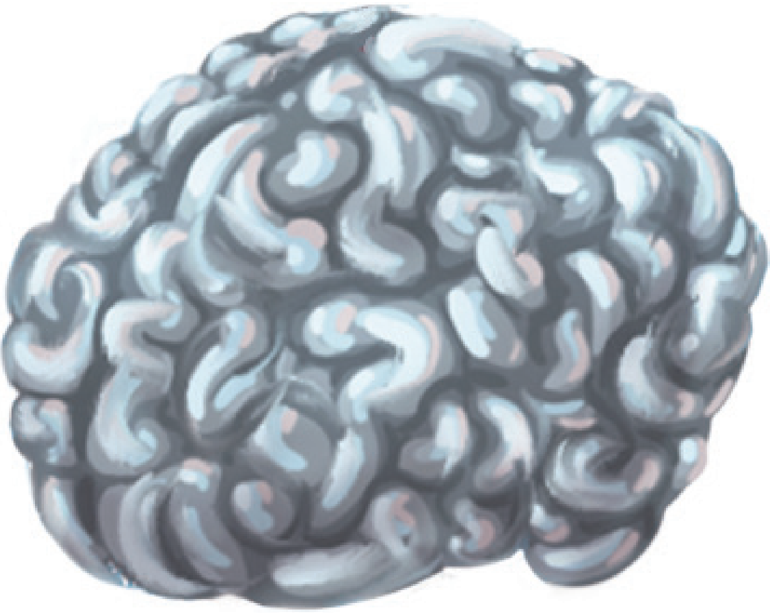
Overall, teacher instruction during the seminar reduced the percentage of lack of knowledge and lucky guesses and increased the number of correct responses, whereas the number of misconceptions was not greatly altered. Based on conceptual change theory and schema theory, we assume that the seminar discussion may not have benefitted the students with misconceptions, as the instructional design to achieve conceptual change should include more explicit activation of students' prior knowledge. A learner's understanding is facilitated by adding new information to an existing mental model or schema in the brain, which comprises relevant prior knowledge a learner has already obtained (Alba & Hasher, 1983; Anderson & Pearson, 1984; Bransford & Johnson, 1972). Cognitive neuroscientists have shown that new information can be added to schemas, i.e. stable neural network, faster when this information fits the prior knowledge (van Kesteren et al., 2012; Tse et al., 2007). Thus, when a learner's schema encompasses a misconception, it may be difficult for the instructor to teach the scientific conception using traditional teaching methods. By activating students' prior knowledge, their current understanding of concepts and potential misconceptions become detectable for instructors (Hailikari et al., 2008; Hewson & Hewson, 1983). Future studies may focus on the use of conceptual change instructions, including activation of prior knowledge, to alleviate the number of misconceptions and enhance students' conceptual understanding.

Our study has several limitations. We focused on the comparison of different measurement instruments for self-perceived knowledge. The relatively low difficulty level of questions might have limited the prevalence of misconceptions as these were more comprehensive than conceptual-oriented questions. Additionally, the post-test questions were not identical but isomorph to the pre-test questions which may have resulted in unanticipated changes in item difficulty because of near-transfer conversions of the original items. Furthermore, students had limited time to process the information as the total length of a seminar was approximately 120min. An additional retention test might have provided more insight into the stableness of students' conceptual understanding.

## Conclusion

This study was conducted to evaluate post-decision wagering as an instrument for measuring students' self-perceived knowledge in educational contexts. Our findings add to the growing number of studies that demonstrate the subjective properties of PDW and thus do not support previous literature on the objectivity of PDW as a measure of confidence. In all, this study supports the use of Likert scales over post-decision wagering as an instrument for measuring self-perceived knowledge in educational settings.

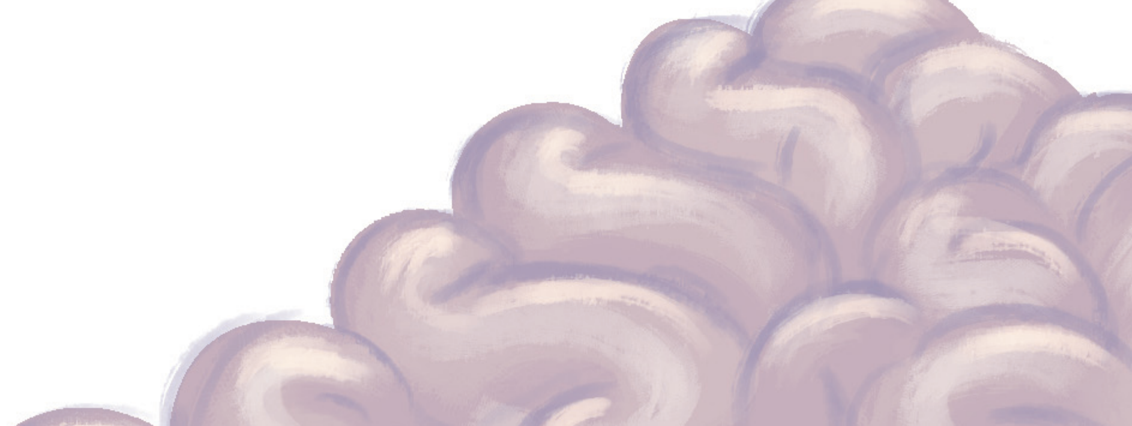
# Chapter 8



# Informing the uninformed: a multitier approach to uncover students' misconceptions on cardiovascular physiology

Marjolein Versteeg · Marjo Wijnen-Meijer · Paul Steendijk

Published in: Advances in Physiology Education (2019), 43(1), 7-14.



## Abstract

Misconceptions about physiology are a major threat to accurate scientific and clinical reasoning in medical education. Awareness is often mentioned as a prerequisite to achieve conceptual understanding; however, students are frequently unaware of their incorrect understanding. We explored the multitier approach as a tool to obtain insight into students' awareness and misconceptions regarding cardiovascular physiology. Biomedical sciences students ($N$ = 81) participated in a diagnostic multitier assessment on cardiovascular physiology. Each question consisted of an answer tier and an explanation tier. Both tiers were paired with confidence tiers, i.e., 5-point Likert scales, which were used as an indicator for metacognitive evaluation, i.e., awareness. The average test score was 3.07 (maximum 4.0) for the answer tier only, and reduced to 1.57 when answer and explanation tiers were combined. A weak correlation ($R^2$ = .13, $p$ = .001) between students' confidence and their test scores was found for the combined responses. Correct combined answers were associated with an increase in confidence score of 0.27 vs. incorrect answers. Using a Bland-Altman analysis, we showed that students generally overestimated their knowledge. In total, 28.7% of all responses were classified as misconceptions, defined as incorrect answers paired with high confidence. In all, findings indicate that the multitier approach is useful to study students' conceptual understanding and uncover misconceptions on cardiovascular physiology. Furthermore, this study supports the need for metacognitive measures in order to improve teaching and learning in medical education.

## Introduction

The awareness of being correct forms an important component of one's knowledge and ability to learn. This idea was already discussed by early and highly influential philosophers, including Aristotle (c. 300 BC) and Confucius (c. 500 BC), and is still acknowledged today (Colman, 2001; Flavell, 1979; Hewson & Hewson, 1983; Legge, 1971). Awareness about one's own thinking and correctness of knowledge after performing a task is here referred to as metacognitive evaluation (Flavell, 1979, Pintrich, 2002). In medical education, students show a lack of metacognitive evaluation skills consistently across medical training: reduced learning outcomes have been demonstrated for prescribing drugs, clinical procedures, and evidence-based medicine (Brinkman et al., 2015; Lai & Teng, 2011; Mavis, 2001; Morgan & Cleave-Hogg, 2002; Morton et al., 2006). Remarkably, little emphasis is put on metacognitive evaluation with regard to conceptual understanding of basic sciences, including medical physiology. Nonetheless, accurate recognition of knowing or not knowing something impacts on students' knowledge acquisition (Friedman et al., 2005; Rangel et al., 2017; Tweed et al., 2017a; Tweed et al., 2017b) and is important for alleviating potential misconceptions in concept learning (Chi et al., 2012; Posner et al., 1982). Any misconceptions on basic science concepts may impair scientific and clinical reasoning, potentially leading to diagnostic errors in medicine (Berner & Graber, 2008; Crosskerry & Norman, 2008; Grande, 2009; Woods, 2007).

Misconceptions are defined as strongly held ideas that are not in line with current scientific views (Chi et al., 2012; Taylor & Kowalski, 2004; Wandersee et al., 1994). Since students have limited metacognitive evaluation skills, we need to develop reliable instruments that can amend this issue (Klymkowsky et al., 2006; Rovick et al., 1999). Making students aware of their level of conceptual understanding may help to alleviate any misconceptions. For instance, research on metacognitive evaluation has shown that students' rating their confidence in their answer could gain deeper insight in their thoughts and potential misbeliefs. Additionally, such confidence ratings may help educators to determine students' actual knowledge (Kampmeyer et al., 2015; Treagust, 1986).

When students remain unaware of their lack of knowledge or misconceptions and subsequently add new information to their current mental structures, this may result in inconsistent thinking (Chi et al., 2012; Posner et al., 1982). Piaget's assimilation theory states that, if there is no fit between the new and the existing information, new knowledge becomes compartmentalised and further strengthens the misconceptions (Piaget, 1976). Educational theorists state that some scientific concepts are difficult to learn because students already hold knowledge that is embedded in naive frameworks, e.g., personal everyday life experiences, and this knowledge is inconsistent with the scientific view (Duit & Treagust, 2012; Duit & Treagust, 1995; Posner et al., 1982). Misconceptions are acknowledged to be highly resistant to change (Chi et al., 2012; Taylor & Kowalski, 2004) and may require educational interventions that differ from the current teaching practice, which only focuses on resolving students' lack of knowledge by providing new information without explicitly addressing misconceptions. Students holding misconceptions may, for instance, benefit from "reshaping" their existing knowledge, also referred to as conceptual change instruction (Hewson & Hewson, 1983), rather than receiving additional factual information or feedback (Archer, 2010; Ecker et al., 2011; Ecker et al., 2010). Thus it is of critical importance to design instruments that allow for measuring students' conceptual understanding, including assessment of their level of awareness via metacognitive evaluation. By implementing such diagnostic tests in the basic sciences curriculum, potential misconceptions can be detected and alleviated accordingly.

One of the most frequently used forms of assessment in medical education are multiple-choice (MC) questions (Anderson, 1979; McCoubrie, 2004). MC questions allow reliable testing of large cohorts and permit evaluation of higher-order problem-solving (Schuwirth & van der Vleuten, 2004), but they are not yet widely applied as instruments for detection of misconceptions. Some studies in medical education research have equipped traditional MC questions with a confidence rating scale (Curtis et al., 2013; Curtis et al., 2012; Grazziotin-Soares et al., 2018; Kampmeyer et al., 2015; Rangel et al., 2017; Tweed et al., 2017a). This allows educators to differentiate between students who are competent, i.e., high confidence-correct, who are guessing, i.e., low confidence-correct, who have a lack of knowledge, i.e., low confidence-incorrect, and who have a potential misconception, i.e., high confidence-incorrect. These MC questions paired with confidence scales are a first attempt toward creating awareness and subsequently uncovering misconceptions by using metacognitive measures in assessments. In the domain of science education, Treagust and colleagues (1998) took the MC assessment a step further (Treagust, 1988; Tsui & Treagust, 2010). They have developed a two-tier diagnostic test to pinpoint students' potential misunderstanding of the subject matter. In the answer tier, students have to make a binary choice (yes/no, higher/lower) about some specific content knowledge. In the explanation tier, students are asked to mark a reason or explanation that supports their choice in the first tier. In the two-tier format, however, it is difficult to distinguish between an accurate understanding or guesswork. Moreover, for incorrect answers, one cannot determine whether the response is a consequence of a lack of knowledge or due to a misconception. As outlined above, the confidence ratings may allow for these classifications. Combining such a two-tier test with additional confidence ratings is also referred to as a multitier approach, which has shown its potential in science education. The multitier approach has been applied in various formats, including the four-tier format, which includes separate confidence scales for each of the two tiers (Caleon & Subramaniam, 2010a, 2010b; Sreenivasulu & Subramaniam, 2014; Yan & Subramaniam, 2018).

The multitier approach is a promising tool to measure conceptual understanding; however, its effectiveness in medical education remains to be explored. The use of a two-tier diagnostic test has been reported in the medical field already, but has not yet been paired with confidence ratings (Palizvan et al., 2013). We aim to investigate if a multitier approach provides information about students' conceptual understanding and potential misconceptions regarding cardiovascular physiology.

The implementation of the multitier approach in medical education may provide insight in students' conceptual understanding and distinguish cases with a potential lack of knowledge from those who hold strong misconceptions. This is useful feedback that can be used by both students and their educators to improve learning and teaching, respectively. Since basic science knowledge forms an important foundation for scientific and clinical reasoning (Berner & Graber, 2008; Crosskerry & Norman, 2008; Grande, 2009; Woods, 2007), we investigate the use of multitier assessments in the context of basic cardiovascular physiology concepts. This research is set out to address the following questions: 1) Can a multitier approach provide information on students' conceptual understanding by assessing their metacognitive evaluation skills? 2) What are the prominent misconceptions regarding cardiovascular physiology and their prevalence among biomedical students?

## Methods

### Participants

Eighty-one biomedical sciences students voluntarily participated in this experimental study. These were second-year bachelor students enrolled in a "Physiology Basic Concepts" course. The male-to-female ratio in this cohort was 30:70.

### Ethical considerations

The educational research study was announced at the beginning of the course, and before the test students were asked to provide informed consent to use their anonymous answers for educational research. They could withdraw their permission at any time. Students received no additional credit, and they were informed that test performance had no effect on their course grade. The study protocol was reviewed and approved by the Leiden University Medical Center Institutional Scientific Committee on Educational Research.

### Study design

As part of the course setup, students were enrolled in seminar groups (~20 students/group). During the seminar, students performed a test consisting of four questions concerning basic cardiovascular physiology concepts: 1) systolic heart failure and ejection fraction; 2) cardiac output and mean arterial blood pressure; 3) transit time in pulmonary and systemic circulations; and 4) afterload and stroke volume. Each question consisted of four tiers, i.e., an answer tier, an explanation tier, and two confidence tiers (Figure 1). In the answer tier, students were asked to provide a binary yes/no (Y/N) answer. In the explanation tier, students could choose one of the suggested explanations (4-6 options) that best supported the reasoning underlying their Y/N answer. Note that each option could be a correct statement in and of itself, but that the students should choose the option that best explains their given Y/N answer. The questions and explanations were designed by a physiology teacher (PS) with longstanding experience in cardiovascular research and teaching and designing and reviewing exam questions. We selected four topics that were handled in the course and on which, based on our experience, misconceptions are relatively common. We aimed for concepts that could be tested by compact statements for which relatively short correct explanations and multiple “plausible” alternative incorrect explanations could be formulated. Confidence was assessed on both the answer and the explanation: “How sure are you that your answer to the previous question was correct?” Confidence was self-graded using a 5-point Likert scale: 1 = very unsure (complete guess), 2 = fairly unsure, 3 = in doubt, 4 = fairly sure, 5 = very sure (almost 100%).

### Outcome measures

To determine whether a multitier approach can make students aware of their conceptual understanding by assessing metacognitive evaluation, we used various outcome measures. First, we reported the performance scores and corresponding confidence scores for each question. Performance scores are also given corrected for guessing, e.g. 25% guess chance for a four-option MCQ, using the following formula: (score − guess score)/(maximum score − guess score). Second, we computed the correlation between overall test performance, i.e., actual knowledge, and average confidence level, i.e., self-perceived knowledge, for each student, using Spearman’s rank correlation coefficient. Third, to determine the within-student difference in confidence between incorrect and correct answers, Question

The transit time of a red blood cell through the pulmonary circulation is less than its transit time through the systemic circulation.

*(1) Answer tier*
Is this statement correct?

A. Yes*
B. No

*(2) Confidence tier*
How sure (confident) are you that your answer to the previous question was correct?

A. Very unsure (complete guess)
B. Fairly unsure
C. In doubt (50/50)
D. Fairly sure
E. Very sure (almost 100%)

*(3) Explanation tier*
Select an explanation for your answer: Because the pulmonary circulation and the systemic circulation are connected in series and ...

A. Pulmonary blood volume is less than systemic blood volume*
B. Pulmonary vascular resistance is less than systemic vascular resistance
C. Right ventricular pressure is lower than left ventricular pressure
D. Blood viscosity is the same in both systems
E. Mean blood velocity is the same in both systems
F. The flow is the same in both systems

*(4) Confidence tier*
How sure (confident) are you that your answer to the previous question was correct?

A. Very unsure (complete guess)
B. Fairly unsure
C. In doubt (50/50)
D. Fairly sure
E. Very sure (almost 100%)

Figure 1 | Example of a multitier question derived from our concept test.

confidence scores were corrected for each student's average confidence score, and subsequently the difference in confidence per question was determined by linear regression.

Following the idea that knowledge is not solely a matter of being able to provide correct answers (performance) but also incorporates students' confidence in it, Kampmeyer and colleagues (2015) introduced the knowledge value. The knowledge value combines correctness and confidence in a single value and allows one to distinguish between learners who have similar test scores, but who differ in their metacognitive evaluation. For example, a student who has the maximum test score but who underestimated her/himself, i.e. a low metacognitive evaluation score, has a lower knowledge value than a student who has both high test and metacognitive evaluation performance. For this analysis, correctness and confidence were both normalised on a 0–1 scale, and we defined knowledge value = 0.5 x (confidence / correctness) and centration = (confidence − correctness) using the Bland-Altman analysis.

We used Hasan's decision matrix to examine the prevalence of misconceptions among students (Hasan et al., 1999). To this end, confidence levels were reduced to dichotomous outcomes: levels 1, 2, and 3 were scored as low confidence; levels 4 and 5 as high confidence. This cut-off was chosen because students selecting "3" were still essentially unsure ('in doubt') about the correct answer. If the student provided the wrong response to the answer tier and indicated that he/she was fairly sure (level 4) or very sure (level 5), a misconception was assumed to be present.

Via the explanation tier, we further tested the students by asking for the underlying arguments. If the student failed to provide the correct explanation, this was also taken to indicate a misconception, even if the Y/N answer was correct.

To identify the misconceptions more specifically, we adopted a quantitative analysis proposed by Caleon and Subramaniam (2010a). To obtain robust results, we classified only incorrect answers and explanations that were chosen by at least 10% of the participants as real alternative conceptions. Accordingly, we calculated the associated confidence in these alternative conceptions (CAC), the average confidence rating of students with this conception. Since we used a cut-off of 3, spurious alternative conceptions in this study have a CAC value between 3 and 3.5 and strong alternative conceptions yield a CAC = 3.5.

## Results

### Metacognitive evaluation

A total of 91.5% students provided a correct response on the answer tier of question 1 (Table 1). Correcting for the 50% guess score, this results in a score of 83.0%. The associated average confidence score was 3.77 out of 5. The explanation tier of question 1 was answered correctly by 54.9% of all students, paired with an average confidence of 3.12. When combining the answers, 54.9% provided a correct answer on both the answer and explanation tier, meaning that almost

Table 1 | Students' correctness and associated confidence on the multitier instrument.

| | Answer tier Y/N | | | Explanation tier | | | Combined tiers | | |
|---|---|---|---|---|---|---|---|---|---|
| | Correct (%) | Confidence | SD | Correct (%) | Confidence | SD | Correct (%) | Confidence | SD |
| Q1 | 91.5 (83.0) | 3.77 | 0.74 | 54.9 (39.9) | 3.16 | 0.84 | 54.9 (48.5) | 3.49 | 0.59 |
| Q2 | 68.3 (36.6) | 3.38 | 0.77 | 26.8 (12.1) | 2.96 | 0.94 | 23.2 (16.2) | 3.17 | 0.67 |
| Q3 | 75.6 (51.2) | 3.67 | 0.85 | 30.5 (16.6) | 3.06 | 0.86 | 29.3 (22.9) | 3.36 | 0.73 |
| Q4 | 70.7 (41.4) | 3.81 | 0.84 | 46.3 (35.6) | 3.48 | 1.01 | 45.1 (40.1) | 3.65 | 0.84 |

*N = 81 students. Scores in brackets are adjusted for guessing.*

40.0% of initially correct students failed to mark the correct explanation. The average confidence for the combined tiers was 3.45. Similar results were obtained for the other questions (Table 1).

The average total test scores (maximum 4 points) among students when combining the answer and explanation tier was 1.57, compared with 3.07 for the answer tier only, yielding a significant reduction in performance ($t(80) = 13.209$, $p = .001$, $d = 1.56$). The self-assessed confidence levels were also significantly lower for the combined tiers (3.42) vs. the answer tier responses (3.67) ($t(80) = 9.337$, $p = .001$, $d = 0.55$). The average confidence levels for the answer tier responses and for the combined responses were both above 3 for all questions.

For the answer tier, there was no significant correlation between students' test

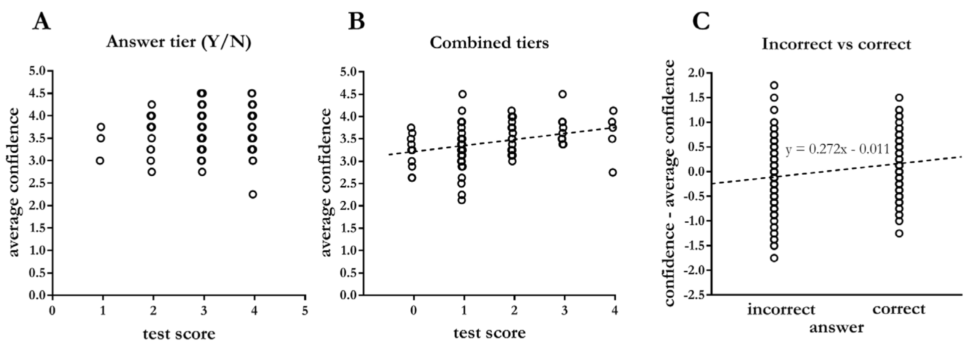



Figure 2 | Relations between confidence and correctness. (A) Average confidence versus test score for the answer tier. (B) Average confidence versus test score for the combined tiers. (C) Regression analysis (within students) shows an increase in confidence related to the correctness of answers. Y/N, yes/no.

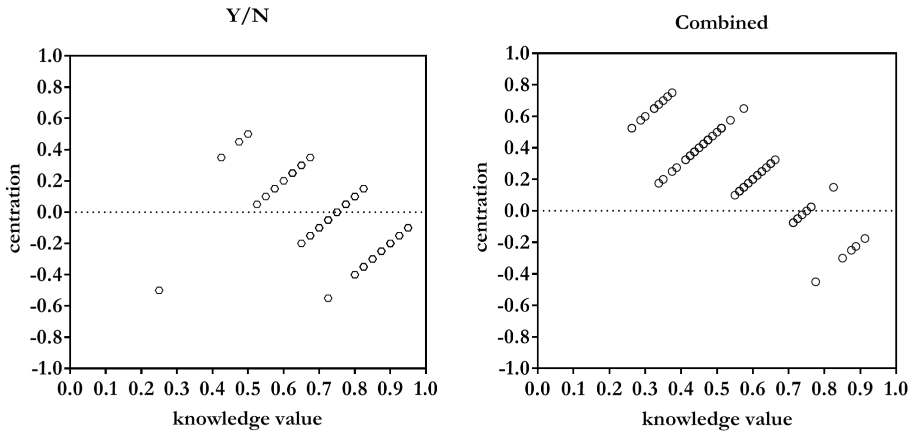



Figure 3 | Bland-Altman analysis. The knowledge values [0.5 x (confidence + correctness)] are plotted against the mean centration values (confidence − correctness) for each individual student. Y/N, yes/no.

scores and their average level of confidence (Figure 2A). For the combined tiers, a weak positive correlation ($R^2 = .13$, $p = .001$) was found (Figure 2B). The average confidence for a correct response on a Y/N question was 3.72 vs. 3.34 for an incorrect response ($t(322) = 1.940$, $p = .053$). For the combined tiers, the average confidence was 3.67 for correct answers and 3.21 for incorrect answers ($t(322) = 2.711$, $p = .007$). After removing the between-student variability in average confidence, the specific effect of an incorrect vs. a correct combined answer was determined as 0.27 on the 5-point confidence scale (Figure 2C). In addition to the correlation analysis, we performed a Bland-Altman analysis to relate students' knowledge values and centration values (compare Methods, Figure 3). The average knowledge value was higher for the answer tier (0.74) compared with the combined tiers (0.54). The mean centration is negative for Y/N answers (0.03), which indicates a bias between actual knowledge and confidence with a tendency toward underestimation for Y/N responses. For the combined tiers, the mean centration is positive (0.29), meaning that, based on their multitier assessment, students, on average, overestimate their actual knowledge.

## Misconceptions

The prevalence of misconceptions was computed using Hasan's decision matrix (Figure 4). Looking only at the answer tiers of each question, 10.4% of all answers were categorised as misconceptions. Almost one-half of the Y/N answers were answered correctly with high confidence (48.8%). Moreover, 27.7% of the answers were correct but paired with low confidence, i.e., lucky guesses, and 13.1% were

Table 2 | Misconceptions on cardiovascular physiology.

| | Misconception | %students with misconception | CAC Mean | CAC SD |
|---|---|---|---|---|
| Q1 | A patient with systolic heart failure will have a low ejection fraction, because with systolic heart failure end-diastolic volume and stroke volume are both decreased. | 34.6 | 3.46 | 0.70 |
| Q2M1 | During exercise the % increase in cardiac output is not approximately the same as the % increase in mean arterial blood pressure. Because during exercise the increase in cardiac output is mainly due to an increase in heart rate. | 30.9 | 3.12 | 0.60 |
| Q2M2 | During exercise the % increase in cardiac output is approximately the same as the % increase in mean arterial blood pressure. Because during exercise the increase in cardiac output is mainly due to an increase in heart rate. | 21.0 | 2.97 | 0.65 |
| Q3M1 | The transit time of a red blood cell through the pulmonary circulation is not less than its transit time through the systemic circulation. Because the pulmonary circulation and the systemic circulation are connected in series and pulmonary vascular resistance is less than systemic vascular resistance. | 28.4 | 3.22 | 1.03 |
| Q3M2 | The transit time of a red blood cell through the pulmonary circulation is less than its transit time through the systemic circulation. Because the pulmonary circulation and the systemic circulation are connected in series and the flow is the same in both systems. | 11.1 | 3.61 | 0.97 |
| Q4M1 | An increase in afterload will generally cause a decrease in stroke volume. Because with an increase in afterload end-systolic volume will decrease. | 16.0 | 4.00 | 0.94 |
| Q4M2 | An increase in afterload will generally not cause a decrease in stroke volume. Because with an increase in afterload end-systolic volume will decrease. | 11.1 | 3.11 | 1.38 |

*CAC, confidence for alternative conceptions.*

recognised as a lack of knowledge.

If also the responses on the explanation tier are taken into account, the distribution of outcomes clearly changes. The percentage of misconceptions nearly triples toward 28.7%, and only 29.9% of all answers are categorised as right conceptions. The percentage of lucky guesses reduces to 8.8%, and lack of knowledge increases to 32.6%.

Table 2 shows a list of seven misconceptions on cardiovascular physiology that were identified in our study. The mean confidence for these misconceptions

Answer tier (Y/N)

| | Incorrect | Correct |
|---|---|---|
| High confidence | 10.4% | 48.8% |
| Low confidence | 13.1% | 27.7% |

Combined tiers

| | Incorrect | Correct |
|---|---|---|
| High confidence | 28.7% | 29.9% |
| Low confidence | 32.6% | 8.8% |

Figure 4 | Hasan's decision matrix. Y/N, yes/no.

ranges between 2.97 and 4.00. Using the classification scheme of Caleon and Subramanian (2010a), we classified two of these responses (Q3M2 and Q4M1) as strong alternative conceptions.

## Discussion

Our study shows that the multitier approach can provide information about students' level of conceptual understanding and their associated metacognitive evaluation skills. Interestingly, students frequently chose an incorrect explanation they believed was associated with their initial correct Y/N response. These incorrect explanations were often paired with high confidence; therefore, we conclude that misconceptions are clearly present among students enrolled in this cardiovascular physiology course.

### Metacognitive evaluation

Based on the Y/N responses, students yielded relatively high performance scores on the test. Students performed significantly worse when they had to mark the explanation they believed was associated with their Y/N response (Table 1). Remarkably, although significantly lower, the self-reported confidence levels were not altered substantially on the multitier assessment compared with the confidence in students' Y/N responses. Average confidence responses were still between "in-doubt" and "fairly sure" and indicated an overestimation of students' knowledge when considering the combined tiers. Furthermore, the relation between students' actual knowledge and confidence showed only a weak correlation (Figure 2). These results are in line with previous literature demonstrating a tendency of students to overestimate themselves (Brinkman et al., 2015; Kruger & Dunning, 1999; Lai & Teng, 2011). Using the Bland-Altman approach, we confirmed that students indeed overestimate their actual knowledge, primarily when they have to choose the right explanation for a question (Figure 3). The negative correlation between the knowledge values and the centration indicated that the overestimation is less for students with a higher knowledge value. These results are in contrast with previous findings from Kampmeyer and colleagues (2015), who found a relatively low percentage of incorrect high-confident answers. However, Kampmeyer et al. (2015) used traditional MC questions instead of a multitier approach, which complicates the interpretation of differences in study outcomes. The difference in confidence responses between correct and incorrect answers is significant in our study, but only for the combined tiers. Our findings are supported by previous studies, which have shown that students' confidence in correct responses is higher (Kampmeyer et al., 2015; Rangel et al., 2017).

**Misconceptions**

The prevalence of misconceptions was 10.4%, considering Y/N responses only, and increased to 28.7% when including the explanation tier (Figure 4). The number of incorrect answers paired with high confidence, i.e., misconceptions, was almost the same as those paired with low confidence i.e., lack of knowledge, indicating that educators should equally focus on both categories. Notably, the percentage of misconceptions in our study was lower than reported previously (Michael et al., 2002; Palizvan et al., 2013), although these studies did not include confidence measures. For cardiovascular physiology, the most prevalent misconceptions of our test were outlined (Table 2). For some questions, the average CAC was higher than for the correct answer. In all, findings indicate that students' basic science knowledge on cardiovascular physiology is insufficient. Therefore, educators should not only design conceptual change interventions to alleviate the prevalence of misconceptions, but also focus on the apparent lack of knowledge among students by examining their prior knowledge, for example.

**Probability of guessing**

The difference between high scores on the answer tier vs. relatively low scores on the combined tiers may be partly explained by the probability of guessing. Since the answer tier is associated with a high chance of guessing (50%), students' scores will be an overestimation of their actual knowledge. The difference between Y/N and combined tiers might thus be more subtle. Therefore, we applied a correction for the probability of guessing to the absolute values (Table 1). Although the difference is smaller after correction for guessing, it remains significant.

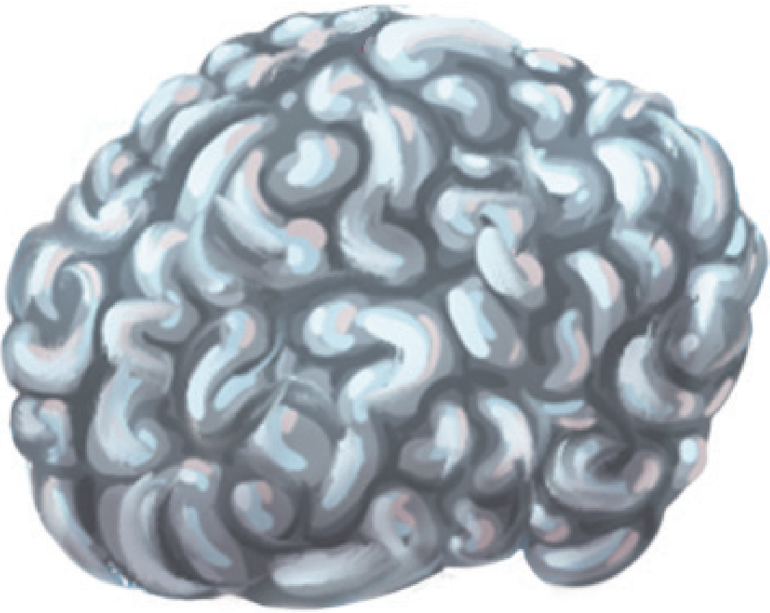
**Strengths and limitations**

To our knowledge we are the first to evaluate the use of the multitier approach in medical education. We used the multitier approach to uncover students' conceptual understanding. Additionally, we showed that there is a clear lack of metacognitive evaluation skills among students regarding basic science knowledge. Still, this study contains several limitations that should be addressed. First, our findings are based on a relatively small set of questions testing conceptual understanding related to cardiovascular physiology. It would be interesting to investigate the multitier approach in other contexts, such as other basic science knowledge or clinical skills. Second, we limited ourselves to exploring the single relationship between students' confidence and performance. Other factors, such as motivation and question type, may have a substantial influence on both confidence ratings and performance and should be analyzed in future studies. Third, we used a 5-point confidence scale for both tiers, although the answer tier only had two answer options. It may not seem logical for students to give a rating 3, as this rating was defined as doubting between two answer options. However, no students commented on this issue, and remarkably 5.9% of the confidence ratings associated with the Y/N response were 3. This finding illustrates that more methodological research might be needed on students' interpretation of confidence rating scales. Lastly, this experiment did not provide a representative overview of existing misconceptions in cardiovascular physiology, since the assessment only comprised four questions. Instead, this study is considered a proof-of-concept study demonstrating that the multitier approach is useful for detecting students misconceptions in medical education.

## Conclusion

We showed that the multitier approach allows students and their educators to gain insight in students' level of conceptual understanding and to reveal their potential misconceptions. Broad implementation of the multitier diagnostic test can help educators to more precisely pin-point knowledge deficiencies, which may result in more effective teaching approaches and learning across the medical curriculum.

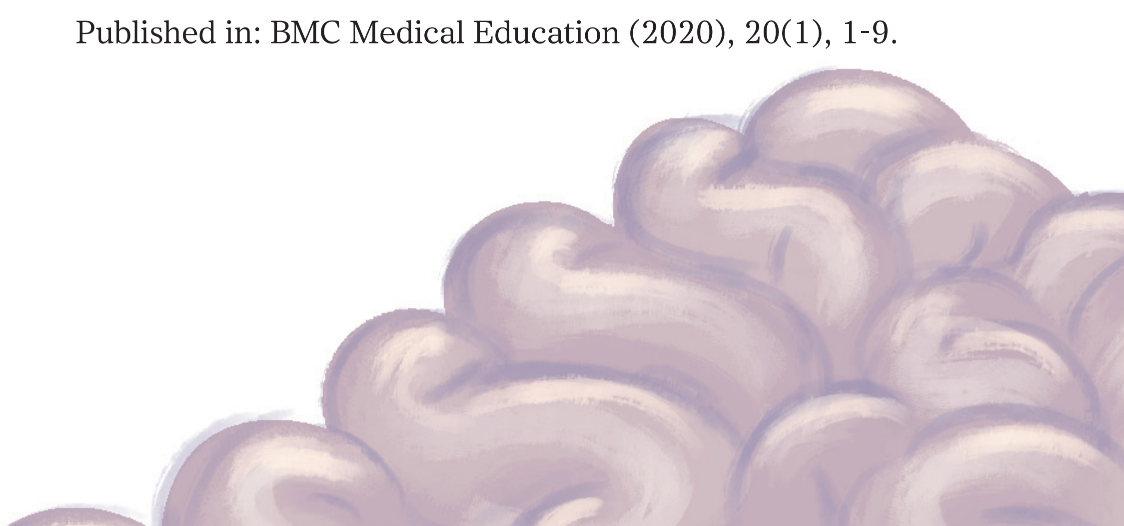
# Chapter 9



# Refuting misconceptions in medical physiology

Marjolein Versteeg · Mariëtte H. van Loon · Marjo Wijnen-Meijer · Paul Steendijk

Published in: BMC Medical Education (2020), 20(1), 1-9.



## Abstract

### Background

In medical physiology, educators and students face a serious challenge termed misconceptions. Misconceptions are incorrect ideas that do not match current scientific views. Accordingly, they have shown to hamper teaching and learning of physiological concepts. Conceptual Change Theory forms the basis of new teaching and learning practices that may alleviate misconceptions and facilitate critical thinking skills that are essential in becoming knowledgeable, self-regulated health professionals. In this study, we examined if such an intervention named refutation texts, could enhance medical students' cognition and metacognition.

### Methods

First-year medical students ($N$ = 161) performed a pre-test and post-test on cardiovascular physiology concepts, including a self-perceived confidence rating. In between, students read either a standard text with an explanation of the correct answer, or a refutation text which additionally refuted related misconceptions.

### Results

In both groups, average performance scores (refutation: +22.5%, standard: +22.8%) and overall confidence ratings (refutation: $\Delta$0.42 out of 5, standard: $\Delta$0.35 out of 5) increased significantly (all $p < .001$), but a significant effect of the specific refutation element was not found. Initially incorrect answers were corrected less frequently in cases of high confidence (35.8%) than low confidence (61.4%).

### Conclusions

Our results showed that refutation texts significantly increased students' knowledge, however, the refutation element did not have a significant additional effect. Furthermore, high confidence in incorrect answers negatively affected the likelihood of correction. These findings provide implications for teaching practices on concept learning, by showing that educators should take into account the key role of metacognition, and the nature of misconceptions.

## Introduction

Physiology plays a central role in understanding human body functions (Finnerty et al., 2010). It is therefore problematic that many medical students find it difficult to acquire accurate knowledge of physiological concepts (Michael, 2007). This may be partially due to the presence of misconceptions (Badenhorst et al., 2016; Michael, 1998; Feltovich et al., 1993). Misconceptions can be defined as incorrect ideas that do not match current scientific views (Versteeg et al., 2019). Over the last 30 years, misconceptions in science education have been demonstrated repeatedly (Ahopelto et al., 2011; Palizvan et al., 2013; Versteeg et al., 2019; Michael et al., 2002). Misconceptions are resistant to change as they frequently persist even after direct instruction (Chi, 2005). In medical education this topic is less well studied and few didactic strategies have been put forward to address misconceptions and promote conceptual change among students.

The process of shifting from an incorrect scientific understanding to a correct one is strongly influenced by what a learner already knows (Bransford, 1979). A learner's prior knowledge should therefore always be engaged while trying to understand new information. Sometimes, the to-be-learned information conflicts with one's prior knowledge. It then requires reorganization of cognitive schemas in the brain to accommodate novel information (Kendeou & O'Brien, 2014; Piaget, 1976). When such cognitive conflicts concern conceptual knowledge, this reorganization process is referred to as conceptual change (Posner et al., 1982). By contaminating the learner's prior knowledge, misconceptions may inhibit rather than facilitate learning of new information (Kendeou & O'Brien, 2014; Sinatra & Broughton, 2011; Guzzetti et al., 1993).

Teachers may not succeed to alleviate misconceptions by simply providing the right answer or explanation to a question (Guzzetti et al., 1993; Tippett, 2010). Educational researchers have proposed that teaching strategies should explicitly undermine, i.e. refute, learners' misconceptions (Braasch et al., 2013), for example by using refutation texts. Refutation texts typically comprise three components: (1) the common misconception, (2) the refutation which explicitly debunks the misconception, (3) the correct answer (Tippett, 2010; Braasch et al., 2013). A refutation text for the misconception that blood slows down at a vessel narrowing states:

> Many people think that the velocity of blood decreases when it enters a constricted section of a vessel, just like cars slow down when the road narrows, i.e. misconception, but this notion is false since blood, being a liquid, cannot be compressed, i.e. refutation. The velocity of blood

actually increases because the same blood volume has to pass through a smaller cross-section in the same time frame, i.e. correct answer.

The potential of refutation texts to induce conceptual change has been demonstrated in various science domains including physics (Braasch et al., 2013) and biology (Mason et al., 2008; Södervik et al., 2014). Superiority of this instructional approach is presumably based on the mechanism of coactivation as described in the Knowledge Revision Components framework (Kendeou & O'Brien, 2013). Coactivation of the misconception and correct concept appears crucial for establishing learners' awareness of the existing conflict. Knowing that there is an apparent inconsistency between prior knowledge and new information may lead to an experience of cognitive conflict, followed by attempts to resolve this issue (Kendeou & Van Den Broek, 2007; Diakidoy et al., 2016). In the case of refutation texts, the cognitive conflict may be induced by reading the common misconception plus a refutation that debunks this misconception. This conflict may lead to the reorganization of cognitive schemas in the brain (Kendeou & O'Brien, 2013; Piaget, 1976). Such reorganizations as induced by refutation texts reflect the process of conceptual change, ideally resulting in accurate conceptual understanding.

The cognitive effect of refutation texts on learners' conceptual understanding depends on a learner's metacognition. Metacognition is the process of thinking about one's thinking (Flavell, 1979), and encompasses an important component referred to as metacognitive knowledge (Pintrich, 2002). Metacognitive knowledge entails being aware of what you do and do not know, e.g. when reading refutation texts this could entail becoming aware of a cognitive conflict (Pintrich, 2002). Only students who can accurately judge, i.e. metacognitively evaluate that their understanding of a concept is insufficient may choose to further study this concept (Mason et al., 2011; Thiede et al., 2003). Building on the importance of metacognition in concept learning, refutation texts may stimulate students' conceptual change through enhancing their metacognitive knowledge (Tippett, 2010; Mason et al., 2011). Various studies in elementary school and higher education support this hypothesis (Mason et al., 2008; Mason et al., 2017), however, others failed to demonstrate enhancement of students' metacognitive knowledge after reading refutation texts (Van Loon et al., 2015). So, despite the theoretical link between conceptual change and metacognitive knowledge, research investigating the influence of refutation texts on learners' metacognitive knowledge remains limited and so far inconclusive.

In addition to investigating the influence of refutation texts on students' conceptual understanding and metacognitive knowledge, it may be of interest to examine which learner characteristics are most productive for facilitating conceptual change. Some studies distinguish between misconceptions held by learners with low-confidence versus misconceptions held with high-confidence (Van Loon et al., 2015; Cordova et al., 2014; Curtis et al., 2013). Of note, in some studies the term misconception is reserved exclusively for the latter type, whereas 'wrong answers' held with low confidence are referred to as lack of knowledge. Researchers have suggested that high-confidence misconceptions are hardest to correct because they are more strongly represented in memory and they impair the student in accommodation contrasting information (Ecker et al., 2011; Pintrich et al., 1993). The hypercorrection effect, however, contradicts this hypothesis, stating that corrective feedback induces coactivation which may particularly surprise learners who are highly confident about their misconceptions, thereby increasing their attention and enhancing text comprehension (Butterfield & Metcalfe, 2001; Butler et al., 2011; Carpenter et al., 2018).

In this study, we investigated the cognitive effect, metacognitive effect and hypercorrection effect in a refutation text intervention. Firstly, we investigated if reading refutation texts improves actual knowledge, i.e. cognition. Secondly, we studied if reading refutation texts improves self-perceived knowledge, i.e. metacognition. Thirdly, we tested if the hypercorrection effect occurred: we hypothesised that high-confidence incorrect answers would be corrected more frequently than low-confidence incorrect answers. In summary, we presume that research on conceptual change interventions in medical education is needed to improve students' conceptual understanding. Moreover, equipping students with accurate knowledge about physiological concepts may ultimately result in improved clinical reasoning and decision-making (Woods, 2007; Lisk et al., 2016).

## Methods

### Participants and setting

This study was conducted in first-year medical students at the Leiden University Medical Center (LUMC). At the start of the academic year the student cohort was divided into 24 groups, 12 groups (12-15 students/group, total 161 students) were included in this study. The protocol was implemented in a seminar on cardiovascular physiology. This seminar was part of a compulsory, 8-weeks course on integrative cardiovascular, respiratory and kidney physiology at the beginning of the second semester. The course seminars focus on solving clinically-based scenarios in small-group sessions led by an expert. This study was performed during the first course seminar, which focusses on the concepts of flow, pressure and resistance. These concepts were introduced and explained in a plenary lecture a few days before the seminar and the students were instructed to study them prior to the seminar also using a specified section from a medical physiology textbook (Boron & Boulpaep, 2016).

This study protocol was approved by the LUMC Educational Research Review Board (ERRB), reference number: OEC/ERRB/20171010/2. Students provided informed consent to use their responses for scientific analysis and publication. They received no additional credit and they were informed that all data would be anonymised and test performance had no effect on their course grade. They could withdraw their permission at any time.

### Procedure

In our study, half of the groups were assigned to the Refutation text intervention, and the other half received a Standard text control intervention. Allocation to these two experimental groups was arbitrarily except for the aim to have a similar male-to-female ratio (30:70) in all groups. The study was performed in a classroom setting at the beginning of a seminar. All students performed a pre-test, followed by either a Refutation text or Standard text intervention, and a subsequent post-test with near-transfer questions. Both tests were given on paper and consisted of four multiple-choice multi-tier questions. All questions were about cardiovascular physiology topics regarding flow, pressure and resistance. In between the pre-test and post-test, each student received either refutation texts or standard texts (see Supplementary L for examples). The standard texts gave, for each question of the pre-test, the right answer plus an explanation (average 177 words/text). Students had four minutes to answer each question on the pre-test and post-test, and also four minutes for reading each refutation or standard text. The study was teacher-paced, meaning that students had to wait for the next question or text if they finished earlier, resulting in a total time of 48 minutes. The refutation texts groups followed a similar procedure, except that the texts contained an additional sentence, i.e. refutation element, that presented a common misconception with an explicit refutation of that misconception, before providing the correct answer with the explanation (average 226 words/text). During the tests, students received a summary sheet with all the relevant factual knowledge to reduce the number of incorrect answers merely due to lack of factual knowledge.

### Materials

The questions and explanations were designed by a physiology teacher (PS) with longstanding experience in cardiovascular research and teaching, and designing and reviewing exam questions. Each question consisted of three tiers, i.e. an answer tier, an explanation tier, and a confidence tier (see Supplementary K for examples). In the answer tier, students were asked to provide a binary Yes/No or an 'Increase/Decrease/No change' answer. In the explanation tier students could choose one of the suggested explanations that best supported their reasoning underlying their answer. Except for the right explanation, all other explanations were designed to reflect possible misconceptions that students may hold. After the explanation tier, students had to answer the confidence tier: 'How sure are you that your answer to the previous question was correct?'. Confidence was self-reported using a 5-pt Likert scale: 1: Very unsure (complete guess), 2: Fairly unsure, 3: In doubt, 4: Fairly sure, 5: Very sure (almost 100%). All questions were designed on the 'apply' and 'analyse' levels of Bloom's taxonomy, and focused on examining students' conceptual knowledge (Krathwohl, 2002).

**Data analyses**

IBM SPSS Statistics Version 23.0 (IBM Corp., Armonk, New York, USA) and GraphPad Prism Version 7.02 (GraphPad Software, La Jolla, California, USA) were used for all data analyses and visualizations. Descriptive statistics are provided as means and standard errors of the mean, unless otherwise mentioned. Only answers that consisted of a correct initial answer and a correct explanation were marked correct. Dependent samples t-tests were performed, for Refutation text and Standard text groups separately, to determine whether there was a difference in pre-test versus post-test scores.

An ANCOVA analysis was used to determine whether the post-test means, adjusted for pre-test scores, differed between groups. To determine the effects of response accuracy (incorrect or correct answer, i.e. cognitive effect), stage (pre- or post-intervention) and group (standard or refutation text), and their interactions with confidence, i.e. metacognitive effect, we used a multiple linear regression (MLR) model with dummy variables. We used effects coding to avoid multicollinearity. Consequently, the coding for the dummy variables for response (R), stage (S) and group (G) was as follows: incorrect answer R=-1, correct answer R=+1, pre-test S=-1, post-test S=+1, standard text G=-1, refutation text G=+1. The MLR model was: $Y = B_0 + B_R.R + B_S.S + B_G.G + B_{RS}.R.S + B_{RG}.R.G + B_{RSG}.R.S.G$.

This model was applied to the individually corrected confidence scores (Y): a student's average confidence score was subtracted from their confidence scores on each question to remove the between-students variability in average confidence scores.

To test the hypercorrection hypothesis, we determined the fraction of initial misconceptions that were changed to a correct answer after intervention and the fraction of initial lack of knowledge that was changed to a correct answer. A hypercorrection effect is found if the fraction corrected misconceptions is

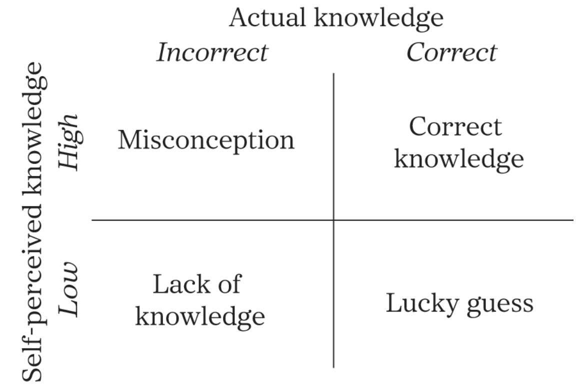



Figure 1 | A decision matrix based on students' actual and self-perceived knowledge. Adapted from Hasan et al., 1991.

higher than the fraction corrected lack of knowledge. Therefore, outcomes were made dichotomous: a confidence score below or equal to 3 was defined as low and a confidence score above 3 as high. This cut-off was chosen because students selecting "3" were still essentially unsure ('in doubt') about being correct. We used Hasan's decision matrix to label the answers (see Figure 1). According to this matrix incorrect answers given with high confidence are considered misconceptions, incorrect answers given with low confidence are considered a lack of knowledge (Hasan et al., 1999). Correct answers held with low confidence were labelled lucky guesses and correct answers with high confidence were labelled correct knowledge. In these terms, misconceptions and lucky guesses are considered low metacognition and correct knowledge and lack of knowledge high metacognition. Furthermore, a cognitive effect was labelled positive when one changed an incorrect answer to a correct answer. A metacognitive effect was labelled positive when one changed from low metacognition to high metacognition.

## Results

Table 1 presents the overall performance and average confidence scores on the pre-test and post-test in both groups. After reading Refutation texts, the overall test performance score increased significantly from 36.3%±0.03 to 58.8%±0.03 ($t(79) = 6.976$, $p < .001$). For the Standard text group a significant increase from 34.3%±0.03 to 57.1%±0.03 was found ($t(80) = 7.198$, $p < .001$). There was no significant difference in post-test performance between the group reading Refutation texts and the group reading Standard texts ($F(1,644) = 0.095$, $p = .758$). The overall confidences scores increased significantly from 3.10±0.06 to 3.52±0.08 in the Refutation text group ($t(79) = 6.154$, $p < .001$). For the Standard text group a significant increase from 3.23±0.06 to 3.58±0.05 was found ($t(80) =$

Table 1 | Percentage correct of overall test performance and average confidence rating.

| Performance | Refutation text ($n$ = 80) | Standard text ($n$ = 81) |
|---|---|---|
| Overall test score, $M$(%) ± $SD$ | | |
| Pre test | 36.3±0.03 | 34.3±0.03 |
| Post test | 58.8±0.03 | 57.1±0.03 |
| Average confidence rating, $M$(max 5) ± $SD$ | | |
| Pre test | 3.10±0.06 | 3.23±0.06 |
| Post test | 3.52±0.08 | 3.58±0.05 |

6.101, $p < .001$). Additionally, there was also no significant difference in post-test confidence scores between the groups ($F(1,160) = 0.003$, $p = .954$).

Figure 2 shows the relationship between students' performance and confidence scores on individual questions. For each student, confidence scores per question were corrected for their average confidence to remove between-student variability in average confidence, i.e. confidence*, see methods. In the refutation text group, the difference in confidence scores between incorrect and correct answers increased from Δ0.419 points pre-intervention to Δ0.643 points post-intervention. The standard text group showed a similar increase from Δ0.382 to Δ0.695 points. A complete overview of the numbers of incorrect and correct answers and related confidence and confidence* in both groups, at both stages is shown in Table 2.

Multiple linear regression analysis was used to determine main and interactive effects of response, stage and group on confidence*, see Table 3. Response and stage were significant predictors. The significant interaction effect between response and stage indicates that the difference in confidence* between incorrect

Table 2 | Number of incorrect and correct answers and related confidence and confidence*.

| Response (R) | Stage (S) | Group (G) | N | Confidence $M \pm SD$ | Confidence* $M \pm SD$ |
|---|---|---|---|---|---|
| Incorrect | Pre test | Standard text | 214 (16.6%) | 3.13 ± 0.84 | -0.13 ± 0.59 |
| | | Refutation text | 197 (15.3%) | 2.92 ± 0.89 | -0.16 ± 0.68 |
| | Post test | Standard text | 141 (10.9%) | 3.14 ± 0.83 | -0.40 ± 0.69 |
| | | Refutation text | 131 (10.2%) | 3.11 ± 0.96 | -0.38 ± 0.61 |
| Correct | Pre test | Standard text | 113 (8.8%) | 3.42 ± 0.99 | 0.25 ± 0.82 |
| | | Refutation text | 116 (9.0%) | 3.40 ± 0.91 | 0.26 ± 0.77 |
| | Post test | Standard text | 189 (14.7%) | 3.91 ± 0.94 | 0.30 ± 0.82 |
| | | Refutation text | 187 (14.5%) | 3.80 ± 0.97 | 0.27 ± 0.74 |

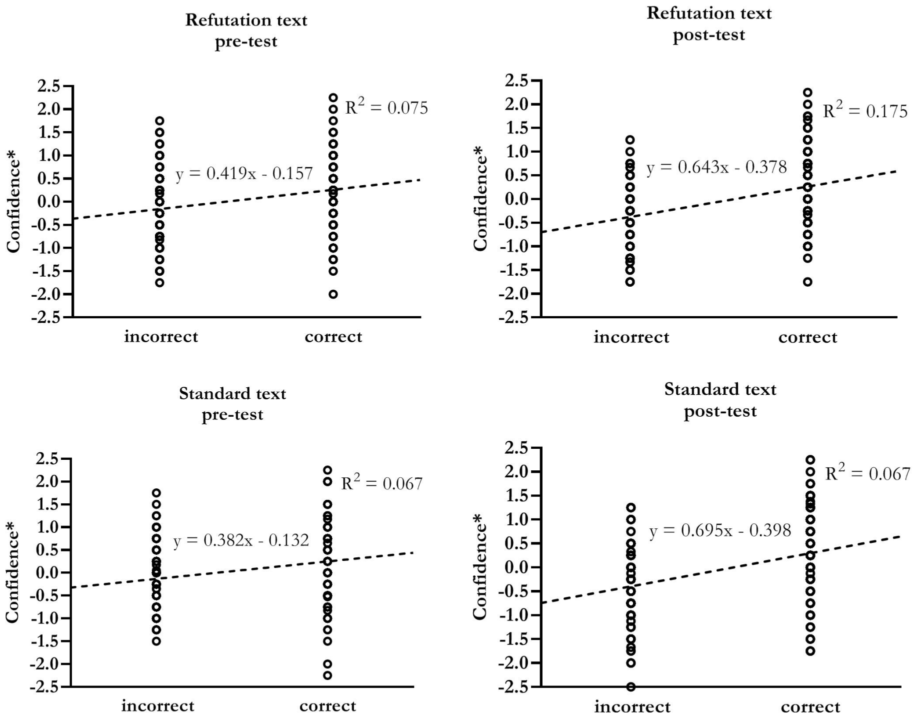



Figure 2 | Relationship between students' performance and confidence*.

and correct answers, i.e. the response effect, was significantly higher post-intervention than pre-intervention. This interaction effect, however, was not significantly different between the groups, as indicated by the lack of significance for the response-stage-group interaction term.

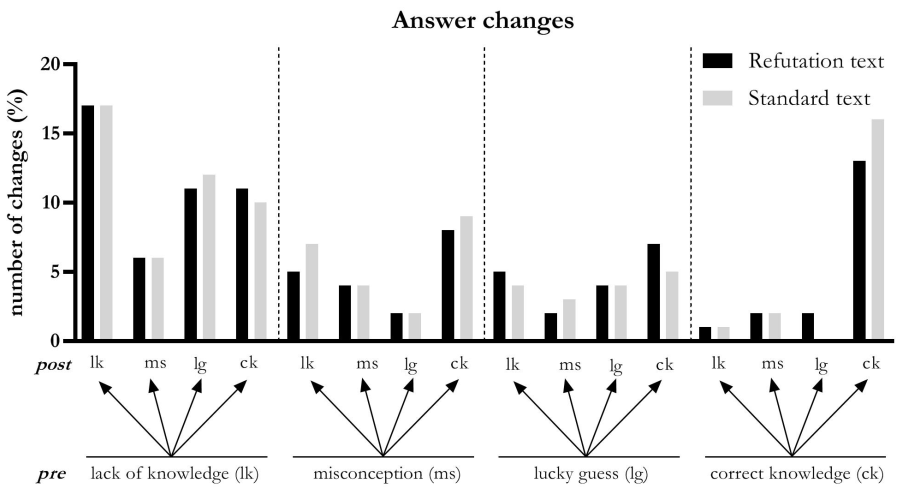
Table 3 | Predictors of confidence*.

| Parameter | B | Significance |
|---|---|---|
| Threshold | 0.001 | .955 |
| Response | 0.267 | .000 |
| Stage | -0.055 | .008 |
| Group | -0.003 | .880 |
| Response x Stage | 0.067 | .001 |
| Response x Group | -0.002 | .930 |
| Stage x Group | 0.00003 | .998 |
| Response x Stage x Group | -0.011 | .585 |

*Response; incorrect answer (-1), correct answer (+1). Stage; pre-test (-1), post-test (+1). Group; standard text (-1), refutation text (+1).*

Figure 3 displays all changes, from pre- to post-intervention, in performance and confidence for the refutation and standard text groups. Both groups showed comparable changes. A positive cognitive effect was indicated if initially incorrect answers, i.e. misconception or lack of knowledge, were changed to correct answers, i.e. lucky guess or correct knowledge. A positive metacognitive effect was indicated if initially low metacognitive accuracy, i.e. misconception or lucky guess, changed to high metacognitive accuracy, i.e. lack of knowledge or correct knowledge. In the refutation text group, an overall positive cognitive effect was measured in 31.9% of cases compared to a negative cognitive effect of 9.4%. The positive metacognitive effect was 23.8% compared to a negative metacognitive effect of 21.3%. In the standard text group, the overall positive and negative cognitive effects were 32.5% and 9.0%, and the metacognitive effects were 25.0% and 20.5% respectively.

The hypercorrection hypothesis was tested by comparing the percentages of low versus high confidence incorrect answers that were changed to correct answers post-intervention. For the Refutation text group, 35.8% of the initially incorrect answers that were rated with high confidence, i.e. misconceptions, changed to correct knowledge after intervention. In contrast, initially incorrect answers rated with low confidence, i.e. lack of knowledge, were corrected to the right conception in 61.0% of cases. Similar findings were obtained in the Standard text group: the percentage of misconceptions that was corrected was 40.3% versus




| | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| cognitive effect | = | = | + | + | = | = | + | + | - | - | = | = | - | - | = | = |
| metacognitive effect | = | - | - | = | + | = | = | + | + | = | = | + | = | - | - | = |

Figure 3 | Answer changes on pre-post test.

66.0% of the lack of knowledge answers. Thus, these data do not support the hypercorrection hypothesis for either group. Rather, students with lack of knowledge more frequently corrected their answers than students with misconceptions.

## Discussion

With this study we investigated if reading refutation texts benefits conceptual understanding in medical students. Based on previous research we expected refutation texts to have a positive effect on students' conceptual understanding, i.e. cognition, plus associated awareness of their understanding, i.e. metacognition. Additionally, based on the hypercorrection hypothesis we suggested that an increase in conceptual understanding would be present in students with high-confidence misconceptions in particular.

In summary, we found that reading refutation texts improved students' cognition and metacognition but these effects were not significantly greater than the effects of reading standard texts. Furthermore, we could not find support for the hypercorrection hypothesis as students' misconceptions were actually found harder to correct than correcting a lack of knowledge. Here, we elaborate on these findings and propose that instructional methods for concept learning should take into account the key role of metacognition, and the nature of misconceptions.

Since the cognitive and metacognitive improvements were found in both groups they could not be attributed to the refutation element, contrasting previous studies in higher education (Braasch et al., 2013; Mason et al., 2008). Rather, the increases may have been caused by the answer and explanation elements that were present in both texts. According to the Knowledge Revision Components framework (Kendeou & O'Brien, 2013), co-activation of learners' prior knowledge and new information may result in awareness of a possible cognitive conflict. Learners attempt to resolve this issue, leading to enhanced conceptual understanding. In our case, co-activation could have been induced by the answer or explanation element rather than the refutation element per se. Notably, many different misconceptions may be present among learners whereas the refutation element only addressed one of the supposedly most common misconceptions. Consequently, other possible alternative conceptions may have been left unaddressed, thereby limiting co-activation.

As indicated by an increase in accurate metacognitive judgements, co-activation seemed to be established to some extent in both groups, although, the absolute metacognitive outcomes remained relatively poor. This finding aligns with a review by Thiede and colleagues (2009) reporting that the average correlation between one's actual performance and one's self-perceived performance across 57 studies was on average 0.27. Since metacognition plays an important role in conceptual change processes, we suggest educators and researchers should pay more attention to the metacognitive component of learning (Sandars & Cleary, 2011). This also relates to the view of the medical education community that students are expected to engage in their education as self-regulated learners (Sandars & Cleary, 2011; Royal College of Physicians and Surgeons of Canada, 2000). Self-regulated learning is an umbrella term that covers the cognitive, metacognitive, behavioural, motivational, and affective aspects of learning (for a review see Panadero, 2017). According to theory and proven in practice, important metacognitive skills to facilitate self-regulated learning are being able to accurately plan, monitor and evaluate one's learning. Being able to self-regulate one's learning is relevant for numerous reasons, including learning effectively, independently, and lifelong (Zimmerman, 1990; Murdoch-Eaton & Whittle, 2012; Bjork et al., 2013). This comes with an important task for the medical educator as explicit teaching of these metacognitive skills inevitable. Learning how to learn cannot be left to students. It must be taught (de Bruin et al., 2017; Tanner, 2012; Gall, 1990).

The lack of additional benefit of the refutation element may be further explained by the nature of our misconceptions. As described by Chi, one can distinguish three types of knowledge representation: single ideas, mental models and categories (Chi, 2009b). Faulty ideas are suggested to be refuted more easily compared to flawed mental models or complex concepts such as physiological concepts. Regarding the latter, learners must generate inferences by connecting and understanding cause-effect relations (Graesser et al., 1994; Kintsch, 1998). For our physiological misconceptions, refutation texts alone may not have been sufficient to achieve coherent concept representation. Additional educational approaches including diagramming, problem-based learning and peer instruction may be needed to establish conceptual change for abstract scientific concepts (Versteeg et al., 2019; Chen et al., 2019; Loyens et al., 2015).

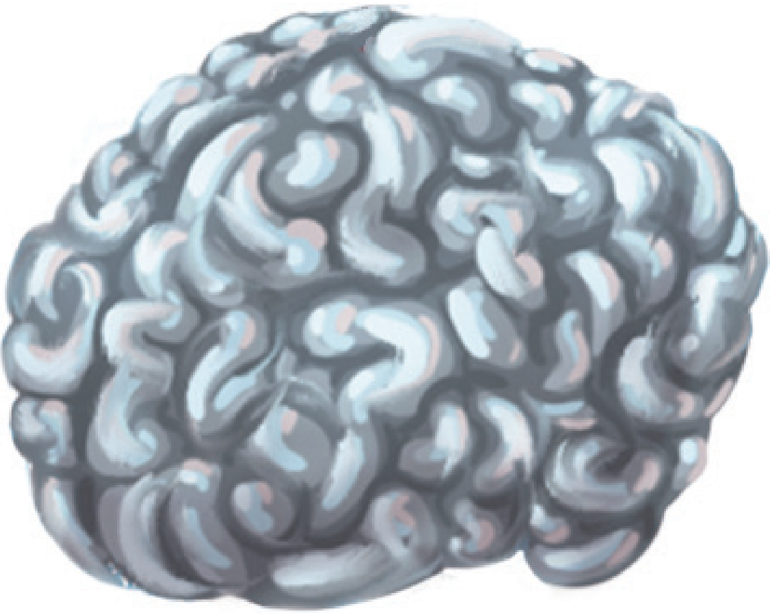
Contrary to the hypercorrection hypothesis, our findings showed that incorrect low confidence answers, i.e. lack of knowledge, were corrected more frequently than incorrect high confidence answers, i.e. misconceptions, after the interventions. This finding suggests that misconceptions are harder to correct than lack of knowledge which resonates with Conceptual Change Theory (Posner et al., 1982). Again, the nature of misconceptions may play an important role in the ease with which conceptual change can be achieved. Interestingly, a previous study by van Loon et al., (2015) showed results similar to our study and suggested that the absence of the hypercorrection effect may also be clarified by the feedback format. Both van Loon et al., (2015) and our study provided feedback (through text reading) to students on all questions, whether they held a misconception or not. Contrastingly, previous hypercorrection studies only provided feedback to learners when they made an error which might cause attentional bias towards the misconceptions (Van Loon et al., 2015). Due to the contextual and protocol variations, it remains difficult to compare and generalise results across studies. Therefore, future studies in the specific context of medicine are needed to advance conceptual change research in medical education.

This research has limitations that need be considered when interpreting its results. Conceptual change is a gradual process, therefore, a longitudinal design including long-term outcomes may provide additional insights in students' learning processes. Furthermore, our study was conducted in a real-life seminar setting and therefore comprised a limited number of questions. Additionally, we used a multi-tier approach with multiple choice answers (Versteeg et al., 2019). Regarding students' cognition, we cannot know if there were other alternative conceptions that were not explicitly stated in the assessment format. Regarding students' metacognition, we cannot identify the metacognitive processes that occurred during reading as we only measured their confidence after reading. Future research may include open-ended questions or thinking aloud procedures to provide more information on students' level of conceptual understanding and metacognitive processes.

## Conclusion

This study was the first to investigate the effect of refutation texts on conceptual understanding in medical students. Reading refutation texts did not significantly improve students' cognition and metacognition beyond reading standard texts. Importantly, we found that misconceptions on cardiovascular physiology were robust and the accuracy of metacognitive judgements among medical students was relatively low. These findings have implications for classroom practice, by addressing the critical role of metacognition and the nature of misconceptions in physiological concept learning. Future studies should take into account these cognitive and metacognitive facets involved in students' learning processes in order to develop effective teaching practices.

# Chapter 10



# What were you thinking? Medical students' metacognition and perceptions of self-regulated learning

Marjolein Versteeg · Guusje Bressers · Marjo Wijnen-Meijer ·

Belinda W. C. Ommering · Arnout J. de Beaufort · Paul Steendijk

Submitted.

# Abstract

### Purpose

A metacognitive habit of mind is essential for healthcare professionals. This study identified metacognitive competencies of medical students as they completed a conceptual learning task, and provided insight in students' perceptions of self-regulated learning in their curriculum.

### Methods

A qualitative study was conducted including a thinking aloud assignment and a semi-structured interview. Eleven third-year medical students from a Dutch University were purposively sampled. The data was transcribed verbatim and analysed iteratively using a template analysis.

### Results

Students differed in their use of metacognitive skills, with an overall focus on monitoring and to a lesser extent on planning and evaluation. Additionally, differences were found in students' metacognitive knowledge and metacognitive experiences. There was apparent use of inefficient superficial predictive cues. Regarding perceptions of self-regulated learning skills, some students felt no need for developing such skills as they perceived medical education as an exercise in memorising facts. Others emphasised the need for more insight in their actual level of knowledge and competence.

### Conclusion

Undergraduate medical students require explicit teaching of metacognitive skills to facilitate self-regulated learning. Educators should aim to integrate metacognition in the everyday discourse of the classroom to foster an environment in which students discuss their own learning.

## Introduction

Self-regulated learning is a necessary skill for healthcare professionals to develop. However, explicit teaching of the required metacognitive competencies for self-regulation appears to be scarce in medical education (Sandars & Cleary, 2011; Artino et al., 2012). Medical students often struggle to acquire an adequate level of metacognition (Lucieer et al., 2016; de Bruin et al., 2017). In this study, we investigated undergraduates' metacognitive competencies and their associated perceptions on self-regulated learning.

Although self-regulation is recognised by the medical education community as an important prerequisite for effective learning, it is a common incorrect assumption that such a skill is implicitly acquired (Bjork et al., 2013). Researchers have suggested that medical schools should specifically emphasise self-regulated learning, since development of self-regulation is a shared responsibility between both students and educators (Sandars & Cleary, 2011; Brydges et al., 2015). Sophisticated self-regulated learners are capable of reflecting on their own performance and modifying their approach for future attempts (Zimmerman, 2000; Sandars, 2013). In essence, self-regulated learners are the captains of their own learning. David Sackett pleads for more captains by advising: "Half of what you'll learn in medical school will be shown to be either dead wrong or out of date within five years of your graduation; the trouble is that nobody can tell you which half—so the most important thing to learn is how to learn on your own." (Smith, 2003). Moreover, not only being able to learn lifelong but also to do it effectively, by setting realistic goals and evaluating one's outcomes, will help students to become better healthcare professionals in the long-term. Multiple studies have reported on benefits associated with adequate self-regulated learning, including enhanced academic achievement, safer and more effective practice, and obtainment of a lifelong learning attitude (Langendyk, 2006; Brydges & Butler, 2012; Zimmerman and Schunk 2013). Self-regulated learning can thus be considered a necessity to all practicing healthcare professionals who have a societal obligation to continuously develop their knowledge.

Self-regulated learning is a cyclical process during which a learner plans one's activities prior to a task, monitors these during a task, and evaluates the outcome after a task (Zimmerman, 2002). The cycle of self-regulated learning is guided by three interrelated components: cognition, metacognition and motivation (Schraw et al., 2006). While most studies in the medical education context focus on cognition (Young et al., 2014) and motivation (ten Cate et al., 2011), research on metacognition is relatively limited.

In the field of social-cognitive sciences, researchers have been studying metacognition for several decades. Flavell and others have described three major components of metacognition, which may inform metacognitive teaching practices in the classroom (Flavell, 1979; Hartman, 2001; Pintrich, 2002; Zohar & Barzilai, 2013). Firstly, one may explicitly teach students about metacognitive knowledge; e.g. educators should help students to make accurate judgements of their learning, so students are aware of what they know and do not know. Secondly, one may explicitly teach metacognitive skills; e.g. educators should provide students with effective strategies for learning so students can use them when studying. Thirdly, one may explicitly teach students about metacognitive experiences; e.g. students' feelings related to the learning task such as a feeling of puzzlement or an aha-experience. Metacognitive experiences are often used by learners as heuristic superficial cues which form the basis for their judgements of learning (Koriat, 1997; de Bruin et al., 2017). For instance, a medical student may judge her/his chances of making the right diagnosis based on the speed with which this diagnosis came to mind. However, fast thinking does not necessarily mean that the student's response is correct. Rather, comprehension-based cues that are formed by causal reasoning are more predictive of correct responses (Thiede et al., 2010).

In undergraduate medical education, we are generally unaware whether and how students use their metacognitive knowledge, skills, and experiences to regulate their learning. Namely, the current focus of research in medical education is on investigating self-regulated learning in the clinical context (Berkhout et al., 2015; van Houten-Schat et al., 2018; Bransen et al., 2019). However, one may argue that metacognition and self-regulated learning should play an important role already at the start of a student's medical training. More research is needed to investigate undergraduates' current level of metacognition and their perceptions on self-regulated learning in the curriculum, in order to develop effective teaching programs.

In this study, we investigated third-year medical students' metacognitive knowledge, skills, and experiences in order to identify the barriers in acquiring an adequate level of metacognition. To put this into perspective, we also asked students about their perceptions of self-regulated learning in the medical curriculum. Ultimately, these findings may support educators in their quest to effectively teach self-regulated learning.

## Methods

### Context

This study was conducted at Leiden University Medical Center (LUMC), the Netherlands. The LUMC offers a six-year medical training program; undergraduate years 1-3 are pre-clinical and graduate years 4-6 consist of clerkships. The Framework for Undergraduate Medical Education in the Netherlands describes the learning outcomes that medical students should achieve in their training to effectively meet the standards of health care. One of the learning outcomes is that undergraduates should possess metacognitive competencies that are necessary to handle a high level of autonomy (Herwaarden et al., 2009). However, in the current LUMC curriculum, formal teaching of such skills is limited. This study was conducted during the third year of the pre-clinical phase. We used this specific population of third-year students as they are expected to have sufficient metacognitive competencies in order to be successful in their clerkships that start after this year.

**Participants**

During January-April 2019, the first author (MV) approached third-year students by e-mail. Participants were purposively sampled to ensure all were in their final year of pre-clinical studies. Additionally, we aimed for a variety in gender that would reflect the medical student population in the Netherlands (30% male, 70% female). A total of 11 students participated in the current study. Due to anonymity we cannot include more potentially identifying information about the participant. Students gave a verbal consent to the audio-recording before the session and signed an informed consent form afterwards. They were given a free lunch worth €7.50 in compensation for their effort. This study was approved by the Institutional Educational Review Board of the LUMC, reference number: OEC/ERRB/20181009/1.

**Data collection**

As little is known about students' metacognitive competencies and their associated perceptions of self-regulated learning in the undergraduate medical curriculum, we designed an explorative qualitative study in which we used a template analysis based on metacognitive theory (Flavell, 1979; Schraw & Moshman, 1995). Accordingly, we conducted a thinking aloud procedure followed by a semi-structured interview to gain insight into students' metacognitive knowledge, skills and experiences. Also, during the interview participants shared their perceptions of self-regulated learning in the medical curriculum. The first author (MV) was present during the thinking aloud sessions and conducted the interviews afterwards.

The thinking aloud assignment and interviews were audio-recorded and transcribed verbatim by the first author (MV). At the start of the assignment, participants were asked to think aloud while solving four exercises on medical physiology. These exercises were designed by an expert physiologist (PS) and aimed to activate participants' scientific reasoning and conceptual thinking. Factual knowledge was presented on an information sheet, so students mainly had to focus on application and integration of the information. After each exercise, students were prompted to evaluate their conclusions by asking how sure they were of their provided answer. If participants were silent for more than three seconds, they were asked to continue to think aloud. Prior to the physiology questions, participants received two practice questions to practice thinking aloud.

The thinking aloud procedure was also used as a prompt for the subsequent semi-structured interview. The interview guide was developed by MV, GB, BO and MW (see Supplementary M). The interview guide was designed and structured based on theoretical concepts of metacognition (Flavell, 1979; Schraw & Moshman, 1995; Pintrich, 2002; Zimmerman, 2002; Zohar & Barzilai, 2013). This guide included questions about goal-setting, learning strategies, and reflective activities. Other issues that were pursued during the interview concerned participants' perceptions of self-regulated learning in the medical curriculum. These questions focused on the value of knowing what you do (not) know and how learning activities could enhance this. Each session, consisting of a thinking aloud procedure and interview, lasted on average 45 minutes. Saturation was reached after eight interviews, after which we conducted three additional interviews to see if saturation was indeed reached.

### Data analysis

Data analysis and collection proceeded in an iterative fashion. The data was coded, analysed and interpreted by MV, GB and MW using template analysis (Brooks et al., 2015), allowing a-priori themes to be used in developing the initial version of a coding template. Metacognitive theory was used to construct the initial template. MV and GB performed open coding on transcripts 1-3 independently. These codes were used to inform the first template. After these three transcripts this process of cross-checking coding, codes and template continued until consensus was reached. As the number of interviews grew, we kept refining the conceptual coding template. The coding template was compared and discussed by MV and GB throughout the data collection period. MV used the template to code transcripts 4-5 and 7-11, continuously refining the template by discussing the findings with GB. Transcript 6 was independently coded by GB. Further refinement of the template through collaborative analysis among the research team led to template consensus (see Supplementary N).

### Reflexivity

The team consisted of researchers with varied backgrounds and expertise in qualitative research to facilitate interpretation of our findings using multiple perspectives. The first author (MV) is a PhD candidate in medical education with a background in neurobiology and has a particular interest in metacognition and conceptual thinking. All other authors are active in the field of medical education research and have different backgrounds, including medical anthropology and sociology (GB), pedagogical sciences (BO), educational sciences (MW), medicine (AJdB) and physics (PS).

## Results

Whereas students were mainly concerned with obtaining a good score, i.e. their cognitive performance, we focused on investigating their metacognitive performance. Firstly, we identified the metacognitive skills that students displayed during problem-solving. With our template we were able to identify if difficulties occurred for specific subtypes of metacognitive skills.

### Planning

Planning occurs prior to the problem-solving process and includes setting goals, selecting appropriate strategies, making predictions, strategy sequences and allocating resources (Schraw & Moshman, 1995). All these themes were found in this study. During the thinking aloud assignment, some students showed planning behaviour by creating an overview of the important characteristics of the exercise, either by highlighting, summarising or visualising. In the interview, students would often recognise their ability to create overview.

> "Well, I think I am good at creating an overview. I will not start problem solving before creating an overview. Generally, I write down all the information and highlight the important information to clarify things before I start problem-solving." (P9)

Furthermore, some students specified that their strategy is to first identify the problem, before doing anything else. However, this was rarely done explicitly during the thinking aloud assignment.

> "Usually, I read the questions first and quickly after that I move on to what is really been asked from me, before reading all the text. However, I did not really do it now [during the exercise]." (P3)

Generally, little time was devoted to planning prior to problem-solving. For instance, explicit goal-setting beforehand by identifying the exact problem and allocation of resources that students would need to solve the question was not observed. Additionally, students would sometimes miss out on information to solve the question because they did not accurately record which information they had at their disposal.

**Monitoring**

Monitoring is the online awareness of comprehension and performance and thus takes place during the actual problem-solving (Schraw & Moshman, 1995). Students used various forms of monitoring, i.e. strategies, during the thinking aloud assignment. Regularly used strategies included rereading, goal-checking, visualising the situation, and eliminating answer options to get to the correct solution. Students rarely switched between different strategies. Generally, they started to use a different approach only when their initial outcome did not align with one of the answer options. Some students admitted that they did not consciously use specific metacognitive strategies during the thinking aloud assignment.

> "I do not really have specific strategies, that I think wow, I should do this or that. I use the strategy to sometimes just read it again. And sometimes you will encounter things during the test that may help you. So yes.. basically like that." (P9)

A large variety in awareness of student's strategy use was found. The excerpt illustrates one end of a spectrum, which encompassed students who did not use any specific strategies at all and were unable to come up with potential strategies they might have used. At the other end of the spectrum were students who were clearly aware of their strengths and weaknesses, and used specific strategies to solve the questions. In the middle of the spectrum, there were students who admitted that, although they knew they should use certain strategies to solve the questions, they did not use them during the task.

The majority of students admitted that they found the questions rather difficult, as they found analytical thinking difficult.

> "I think we are trained in medical school to learn factual knowledge, and this [exercise] is a different skill than learning facts, or connecting facts, so... this is really a different skill, so I think that is always difficult, but it requires quite some brainpower." (P2)

In general, students felt that analytical reasoning is a competency which is not actively taught during medical training.

**Evaluation**

Evaluation refers to appraising the outcomes and regulatory processes after problem-solving. This includes, for example, evaluating one's goals and conclusions (Schraw & Moshman, 1995). During the thinking aloud assignment, few students evaluated their goals and conclusions after marking one of the answer options. We prompted the students to evaluate by asking how sure they were of their provided answer. Despite this prompt, students spent little time evaluating their answers. Most students would answer the question 'How confident are you that your given answer is correct?' without explicitly elaborating on their feeling of confidence or without checking their answer. During the interviews, most students described that a 'feeling of logic' would determine their level of comprehension.

> "Yes, if my feelings tell me that it [the answer] is not right but according to the formula it would be right, then I think, this is not right so I will doubt. Like, when it is not in accordance with each other, and if I cannot solve it with the formula and with my feelings, then I am not sure." (P6)

Other cues for comprehension included; time spent on task, familiarity with the learning material, and the ease of reasoning.

> "I always notice that the longer I think about it, the more I start doubting." (P1)

> "Yes, but you also think I really have known this [learning material]. I have really studied this and known this. I didn't know then if I was good at it, but yeah." (P2)

> "I am sure when; this is how I reasoned and then I get to the right answer, then this seems the right answer to me." (P5)

Interestingly, the feeling of logic and familiarity with the learning material were often mentioned, whereas these cues generally are not good predictors for the level of comprehension (Thiede et al., 2010; de Bruin et al., 2017). Rather, the ease of reasoning is well associated with one's actual level of comprehension but this cue was mentioned to a much lesser extent.

**Perceptions on self-regulated learning in the medical curriculum**

During the interviews, the thinking aloud data was enriched by asking students to reflect on the role of 'knowing what you (do not) know' in the curriculum. Students said they valued the ability to accurately estimate their knowledge and skills so they know what they are (not yet) competent at. Most students outsourced this ability to external assessment tools. For example, they would mention both formative and summative exams, study assignments, and e-learnings as tools to estimate their level of knowledge or skills.

> "Yes, you have mostly study assignments and practice exams of course which have a diagnostic value in terms of what knowledge you actually already possess." (P8)

A few students described the ability to accurately estimate one's knowledge as an internal, personal ability that could be developed by specific strategies, such as self-explaining, explaining to a peer, and consulting a teacher or other sources. The word 'reflection' was rarely mentioned, and if it was, students indicated that focus on this competency during their medical training was insufficient.

> "But at such a meta level not really I think. That we really reflect on ourselves in terms of how well we understand something? How well do we understand exercises, or how well do we understand how we have to handle knowledge and things. I think that should be an essential component of an educational program, especially of an academic education.. There is not enough attention paid to it [reflection] I think. We have to write reflection reports but you might as well just fill in three words, because the teacher is OK with it anyway. Yes, it is mainly just a fill in assignment, and not really that you, as a student, will take a look at yourself thinking what can I do better. And if you would take that maybe a little bit more seriously, also looking [as a teacher] what he [the student] actually does with it [the feedback], because that is being forgotten most of the time." (P8)

Students are in search of 'hard outcome measures', i.e. numbers, as they mention that they would appreciate having more insight into the status of their competences.

> "Uh... with those competences, so competency-based education, really the practice-based education, I would have to admit that I find it really difficult to monitor my progress and how I have developed myself. That

> of course is one of the subjects during the teacher-coach conversations, but it is not very tangible. How good of a communicator, how good of a team player have I become during the last three years? I do not really have a clue, and I do not have any numbers either." (P7)

All students offered ideas in response to one of the questions on how we could enhance one's insight into one's learning during the medical curriculum. Students mostly mentioned that having more of the currently existing assessment tools, e.g. low-stakes exams, study assignments and e-learnings, would provide more insight in one's knowledge deficiencies.

> "But maybe if, for example, the self-study assignments or e-learning assignment are designed like the exam, it would be.. but, then your learning is very exam-oriented maybe." (P2)

A few students elaborated on the benefits of intensifying feedback and reflection to facilitate personal continuous development.

> "Well, actually we were talking about this yesterday in an educational committee. Longitudinal assessment and improving yourself and such… that more attention should be paid to that. And that you maybe can ask the students themselves or assess if.. are they willing to improve themselves?" (P8)

Finally, some students mentioned that medical education is solely about learning medical facts.

> "To me, what I am learning is mostly about learning facts, and that is what's assessed really… I am not looking for any help [in learning facts], because it's things that I know or do not know. I don't think that anyone can help me to better learn things by heart, because I, that is one of my strengths, that I am good at learning things by heart. So, I don't really need help with that and I think that during my education that [learning facts] is what I mainly do for an exam, and in that sense, to a lesser extent understanding the material." (P7)

As illustrated by this excerpt, students felt that learning facts does not require any monitoring, e.g. help-seeking, indicating that developing metacognitive skills may not be necessary to succeed in medical school.

## Discussion

Our study provided insight in medical students' metacognitive knowledge, skills and experiences. Additionally, we obtained students' perceptions on self-regulated learning in the medical curriculum.

Regarding metacognitive skills, students used various ways to monitor their learning process while problem-solving. For example, they visualised the situation to make the problem less abstract. Contrary to monitoring, less time was spent on planning and evaluating. Previous research on metacognitive skills in clinical reasoning also showed that students performed monitoring, but that planning occurred to a much lesser extent (Artino et al., 2014). Planning and evaluation are strong predictors of academic performance (Murad et al., 2010; Patel et al., 2015; Gandomkar et al., 2016). Importantly, these skills are modifiable and teachable, rather than fixed traits (Zimmerman, 2000). Tanner has provided examples of self-questions that learners may ask in training their metacognitive skills, either on the level of an assignment, a single class session, an exam, or a full course (Tanner, 2012). These questions are not only helpful for learners, but also serve as a tool for educators who aim to address metacognitive skills explicitly in their classrooms.

Regarding metacognitive knowledge, a large variety in awareness about one's learning process was found. For those students with little metacognitive knowledge about types of skills or how to use them, there is a need to teach this explicitly (Pintrich, 2002; Tanner, 2012). This includes teaching various strategies, various cognitive tasks, and accurate knowledge about themselves. Moreover, educators should be aware of students' prior knowledge about a subject before teaching them new information. For example, preassessments may be very valuable tools in encouraging students to examine their level of knowledge, and for educators as a diagnostic tool to gain insight in students' understanding (Versteeg et al., 2019). Educators should take responsibility, especially since we know from literature that students themselves are rather poor judges of their actual knowledge and competencies (Thiede et al., 2003; Versteeg et al., 2019).

Regarding metacognitive experiences, most students estimated their performance based on a feeling of logic or a feeling of familiarity, and to a lesser extent on the ease of reasoning. The first two can be referred to as surface-related cues that operate automatically and unconsciously, and which are generally unreliable as predictive cues for performance (Koriat, 1997; Thiede et al., 2010). Importantly, learners can be trained to effectively use predictive cues such as comprehension-based cues, e.g. ease of reasoning (Begg et al., 1989; Koriat, 1997). Various examples of training methods include, generating key words or summaries in case of learning factual knowledge, and completing diagrams in case of conceptual knowledge (Thiede & Anderson, 2003; van Loon et al., 2014). Teaching students explicitly to recognise and generate predictive cues in the classroom may eventually lead to enhancing predictive cue use during self-regulated learning outside of the classroom (de Bruin et al., 2017).

**Facing the facts**

During their medical training, students continuously have to prove themselves, resulting in their learning being driven by assessments (Wormald et al., 2009; Boulet & Durning, 2019). Our research confirms this as students describe the value they assign to assessment outcomes. They feel that assessments are the main indicators for performance. Regarding competency-based education specifically, students felt they have no insight in their level of competency as they cannot fall back on any numerical indicator of performance. The impact of assessments on medical students' motivation to study is profound and often leads to a surface approach to learning (Marton & S□lj□, 1976; Wormald et al., 2009; Cilliers, 2015). This surface approach is characterised, for example, by students' aim to memorise facts (Ramsden, 2003). Notably, our students underlined this statement by describing that the focus in undergraduate medical training is on learning factual knowledge. They felt this approach to learning comes at the cost of their reasoning abilities. Even clinical 'reasoning' is described by some students as a process of pattern recognition during which one has to merely recall factual knowledge. The medical education community has already suggested that one of the solutions to establish deep-learning might entail more integration of basic science and clinical learning, which would meet the students' needs for conceptual knowledge to better understand medical concepts (Kulasegaram et al., 2013). Additionally, since assessment drives learning, educators and faculty should better align assessments with the skill sets required for practice. Currently, there is a trend towards integrated longitudinal assessment programmes that facilitate a more continuous evaluation of student abilities, and which aim to produce competent lifelong learners. Moreover, we may consider rewarding students' metacognitive performances, such as their use of reflective journals (Tanner, 2012).

> "The point at which students have both learned metacognitive skills and have become aware of when to apply these strategies is hypothetically the point at which they have matured into lifelong learners within their disciplines." (Tanner, 2012).

Defining the necessary competencies allows for the integration of meaningful assessments in medical education that help our students to become lifelong learning health professionals (Schuwirth & van der Vleuten, 2012; Boulet & Durning, 2019).

**Strengths and limitations**

A strength of our study lies in combining the thinking aloud assignment with an interview session. This approach allowed us to better grasp all three facets of students' metacognitive performance. It also functioned well as a prompt for students' thoughts about self-regulated learning on the curriculum level.

Some limitations must be taken into consideration. First, the study was conducted in a non-authentic setting, meaning that contextual factors from a real environment which may influence learning behaviour and performance were excluded. Second, metacognitive skills were measured without explicitly asking our students about these skills during the task. This is contrary to the use of microanalyses which are defined as "structured interview approaches that involve administering context-specific questions targeting multiple cyclical phase processes as trainees engage in authentic activities" (Cleary et al., 2016). These microanalyses are often used to measure self-regulated learning processes and prompt students to focus on strategic steps during problem-solving (Artino et al., 2014). However, such prompts may trigger students' awareness and induce 'artificial' use of metacognitive skills as they may not have used these skills in a non-prompted setting. Therefore, we feel that our approach leads to a more accurate image of students' use of metacognitive skills.

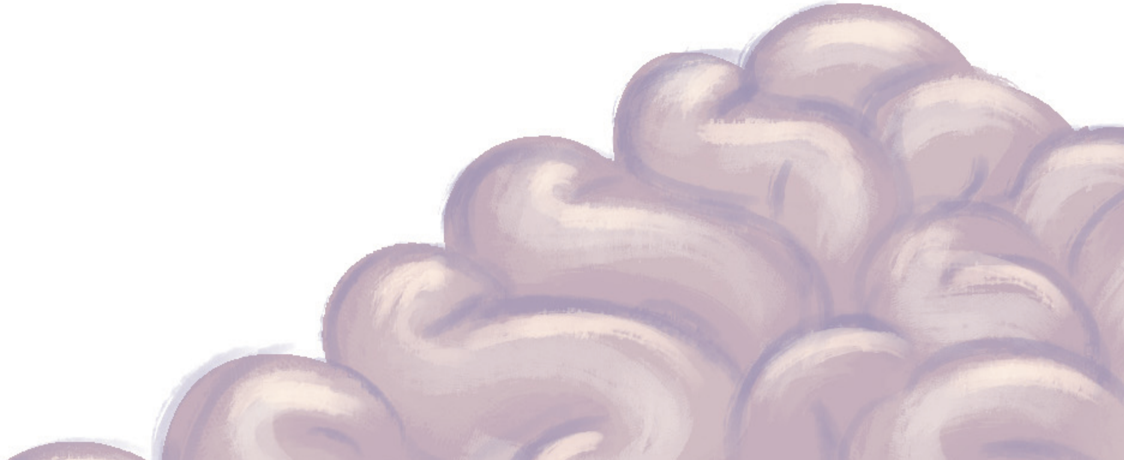
Our study was performed among final year pre-clinical medical students in one Dutch University. Despite this specific context, we feel that our findings are transferable within the Dutch educational context as all Dutch medical programs are based on the same blueprint as developed by the Dutch Federation of University Medical Centers (Herwaarden et al., 2009). Additionally, part of the context is transferable to the medical education community as a whole, since medical education programs worldwide consist of a preclinical and clinical phase.

## Conclusion

This study revealed that medical students are in need of explicit training of metacognitive skills to facilitate self-regulated learning. This includes planning, monitoring and evaluation skills. Moreover, our findings showed that the level of metacognitive knowledge and metacognitive experiences highly varied among students. Educators and faculty should aim to integrate metacognition in the everyday discourse of the classroom to foster an environment in which students discuss their own cognition and learning. This includes the use of novel assessment strategies that drive both cognitive and metacognitive learning in order to develop metacognitive habits of mind and stimulate lifelong learning.

# Chapter 11

# General discussion



## · Purpose ·

Researchers and policymakers have acknowledged the need for evidence-informed curricula to support effective health professions education. Linking brain research to education research contributes to the development of such curricula by informing us on the science of learning. Namely, the brain influences learning, and learning influences the brain. This interplay between the brain and learning, also referred to as educational neuroscience, lies at the heart of the studies in this thesis.

The educational neuroscience-inspired research in this thesis focuses on three specific learning processes: spaced learning, concept learning, and metacognitive learning. Together, these learning processes are considered relevant as they are closely related to the fundaments of Bloom's Taxonomy for establishing educational goals and to the educational goals of health professions' curricula.

The overall aim of this thesis is to improve health professions education by investigating spaced learning, concept learning and metacognitive learning using an educational neuroscience-inspired approach. This general discussion chapter elaborates on how this aim has been fulfilled. First, I provide the main findings for each learning process, including the implications for research and practice. Second, I provide my reflections on the methodology. Third, I discuss the way forward by addressing the potential of educational neuroscience-inspired research.

## · Spaced learning: Finding the right time to learn ·

*"Learning that is easy is like writing in sand, here today and gone tomorrow." - P. Brown.*

By performing a scoping review, we discovered that spaced learning is finding its way to health professions education, although this mainly occurred over the last ten years **(Chapter 2)**. This implies that the spacing effect, which dates back to 1885, needed more than a hundred years to transfer from cognitive psychology in order to make a substantial contribution to the health professions education literature (Ebbinghaus, 1885). Recent implementation might be a result of an increasing interest in online learning tools by early adopters (Rogers, 2003). These early adopters value the trialability attribute of innovations, resulting in a focus on improving educational practices and less on advancing theory or knowledge. Indeed, the majority of spaced learning literature in our community focuses on online settings. Perhaps this idea about early adopters also explains our finding that current definitions and applications of spaced learning often lack the necessary detail to support implementation or replication. Consequently, we suggest that researchers - including early adopters - should base their studies on previous empirical research or theory to facilitate implementation and replication.

As an example, we designed a study in which the spacing format was based on neuroscientific theory **(Chapter 3)**. Although our theory-informed spacing format did not enhance knowledge retention in medical students, we provided an informed and detailed study design that can still be used as a to-be-adjusted template by future researchers. Additionally, we implemented our format within a lecture setting, whereas most spaced learning research is conducted in online settings and simulation training. Our study may encourage researchers and educators to experiment with spaced learning in different educational settings, such as lectures or working groups.

We also learned an important lesson regarding the translation of a neuroscientific hypothesis to an educational environment: to carefully take into account the complexity of the environment. For instance, there are various ways to incorporate units of repetition in an educational setting, e.g. summaries or quizzes, whereas neuroscientific experiments always use an identical repetition mode, i.e. “high-frequency stimuli” (Smolen et al., 2016). Moreover, we used summaries as our repetition mode which can be classified as passive repetition, whereas the positive effect of spaced learning on knowledge retention may occur only when repetition is more actively practiced, such as with quizzes (Roediger & Butler, 2011).

To conclude, spaced learning research is need of clearly described applications to facilitate implementation and replication. While studying these applications, one should take into account - and be specific about - the characteristics of the study environment in order to determine if findings can be generalizable across educational contexts.

**Practice points for researchers**

› Use a clear definition of spaced learning in describing your research in order to stimulate consistent application.

› Describe the study design thoroughly, using empirical research and/or theory.

› Experiment with spacing formats using different interstudy intervals, retention gaps, and forms of (active) repetition.

**Practice points for educators and policymakers**

› Further expand on the use of spaced learning in online settings and simulation settings, aiming to find optimal spacing formats.

› Start implementing spaced learning in different settings such as lectures or working groups, and evaluate the effects in your educational context.

## · Concept learning: Know your misconceptions ·

*"Nothing is more dangerous for a new truth than an old misconception."- J. von Goethe.*

We demonstrated that medical students have various misconceptions that hamper their understanding of physiological cardiovascular concepts such as 'pressure', 'flow', and 'resistance' **(Chapter 4)**. Misconceptions are robust incorrect scientific beliefs that are difficult to alleviate as learners are generally unaware of their own misconceptions even after taking corresponding university classes (Chi et al., 1994a; Palizvan et al., 2013). With the help of educational theories on concept learning, we were able to categorise the misconceptions found in our study. This categorization may serve as a tool for educators to more precisely pinpoint misconceptions in their students and develop more effective instructions accordingly.

To know what such an instruction should entail, we investigated the underlying mechanism of (mis)understanding scientific concepts **(Chapter 5)**. By measuring students' brain activity using functional magnetic resonance imaging, we found significant activation of the putamen in misunderstanders (holding a misconception) compared with understanders (holding the scientific conception). This finding suggests a role for episodic memory in students holding a misconception. Although it is a long way from brain imaging to the development of instructional designs for educational practice (Howard-Jones et al., 2016), our findings may provide directions for concept learning in health professions education and other scientific disciplines. Based on the brain activity we found in the putamen, we support previous recommendations that focus on the crucial role of prior knowledge in teaching scientific concepts (Hewson & Hewson, 1983). Namely, a learner's understanding is facilitated by adding new information to an existing mental model or schema in the brain, which comprises relevant prior knowledge a learner has already obtained (Alba & Hasher, 1983). Cognitive neuroscientists have shown that new information can be added to schemas, i.e. stable neural networks, faster when this information fits the prior knowledge (Van Kesteren et al., 2012). Thus, when a learner's schema encompasses a misconception, it may be difficult for educators to teach the scientific conception using traditional teaching methods. Instructional designs for concept learning should therefore aim to disclose students' prior knowledge and assumptions about the concept of interest.

An instructional design named peer instruction has demonstrated to facilitate conceptual understanding, potentially by disclosing students' prior knowledge **(Chapter 6)**. Peer instruction is a method that includes questions and subsequent discussion among peers (Mazur, 1997). The questions may activate students' prior knowledge which they then make explicit to each other during the peer discussion. This mechanism may be effective not only for peer instruction, but also for other forms of interactive learning (Chi & Wylie, 2014). An interesting finding derived from this study is that even two initially incorrect peers can get to a correct answer together. This finding is in accordance with current literature and supports the idea of making one's thoughts explicit to another learner or the teacher, even if these thoughts are incorrect (Smith et al., 2009).

To conclude, students often have an inadequate level of conceptual understanding and educators may pinpoint students' misconceptions more easily by using categories. Additionally, both students and educators should acknowledge the importance of prior knowledge, and make this explicit, for example through interactive strategies. For peer instruction specifically, future research may take a qualitative approach which can provide additional insight in the reasoning processes of students and other valuable aspects of interactive instructional designs.

**Practice points for researchers**

› Investigate the origins of misconceptions and misunderstandings in your discipline to direct educators towards the challenges in concept learning.

› Explore the underlying mechanisms of concept learning to advance knowledge on how conceptual understanding comes about.

› Use qualitative approaches to uncover the working mechanisms of interactive learning strategies and the role of prior knowledge in concept learning.

**Practice points for educators and policymakers**

› Identify misunderstandings and misconceptions in your teaching - either empirically or by using literature/databases - and make them explicit.

› Aim to disclose students' prior knowledge and assumptions.

› Try interactive strategies such as peer instruction to improve students' conceptual understanding.

## · Metacognitive learning: Comfortably confident? ·

*“Learning how to learn cannot be left to students. It must be taught.” - M. Gall*

Across multiple studies, we found that students show difficulty estimating what they do (not) know, regarding conceptual knowledge specifically **(Chapter 7, 8, 9).** The ability to estimate one’s knowledge after performing a task or test is referred to as metacognitive evaluation (Zohar & Barzilai, 2013). We found that the multitier approach is an effective tool to obtain insight into students’ metacognitive evaluation ability **(Chapter 8)**. Students took a multitier assessment in which a multiple choice question is paired with a confidence tier, i.e. 5-point Likert-scale **(Chapter 7)**, as an indicator for students’ self-perceived knowledge. Although it is difficult to translate confidence to a numerical value, assessing confidence via the multitier approach appears an easy and effective way to diagnose students’ knowledge deficiencies. This is informative for both students and educators, since knowing that the level of conceptual understanding is insufficient is necessary for choosing to further study or teach this concept.

Besides a diagnostic tool, we also developed refutation texts as an intervention aimed at enhancing students’ metacognitive evaluation ability, and conceptual understanding accordingly **(Chapter 9)**. A refutation text undermines a learner’s misconception by explicitly debunking it (Tippett, 2010). Although the refutation texts in our study did not significantly improve students’ cognition and metacognition, they provide an example of an intervention that is focused on enhancing metacognitive learning. Such interventions directed at improving metacognitive competencies are important to develop in health professions education, as they play a crucial part in the path towards an accurate conceptual understanding.

The studies in Chapters 7, 8 and 9 all raised similar questions: Why are students’ estimations of what they do (not) know inaccurate? Can this be attributed to inadequate metacognitive competencies additional to metacognitive evaluation? To answer these questions, we qualitatively explored the metacognitive competencies of medical students as they completed a conceptual learning task **(Chapter 10)**. We found that students differed in their use of metacognitive skills, with an overall focus on monitoring and to a lesser extent on planning and evaluation. Although self-regulation, including metacognitive learning, is recognised by the medical education community as an important prerequisite for effective learning (Sandars & Cleary, 2011), it is a common incorrect assumption that such a skill is implicitly acquired (Bjork et al., 2013). Researchers have suggested that medical schools should specifically emphasise self-regulated learning, since development of self-regulation is a shared responsibility between both students and educators (Sandars

& Cleary, 2011). Based on our findings, we suggest they may focus on the use of planning and evaluation skills of students. Tanner (2012) has provided examples of self-questions that learners may ask in training their metacognitive skills, either on the level of an assignment, a single class session, an exam, or a full course. These questions are not only helpful for students, but also serve as a tool for educators who aim to address metacognitive skills explicitly in their classrooms.

It was worrisome to hear that some medical students felt no need for developing such metacognitive skills as they perceived medical education as an exercise in memorising facts. This remark calls for a medical curriculum reform in which more emphasis is laid on understanding, applying, analyzing, evaluating, and creating knowledge instead of remembering knowledge. Such a reform may enhance medical students' higher-order thinking skills and stimulate a lifelong attitude among our future doctors.

To conclude, educators may use various methods, e.g. multitier assessments and refutation texts, to facilitate (their insight in) students' metacognitive learning. Taking into account the metacognitive facets of learning - and teaching them to students explicitly - may help to enhance students' understanding of the subject matter. For our future medical doctors specifically, we may consider an undergraduate curriculum reform in order to facilitate a lifelong learning attitude.

**Practice points for researchers**

› Take into account the metacognitive facets of learning when investigating conceptual understanding.

› Gain insight in students' metacognitive competencies and development thereof in their medical training, especially in the undergraduate curriculum.

**Practice points for educators and policymakers**

› Implement a multitier diagnostic test to gain more insight in students' level of knowledge and to pinpoint misconceptions.

› Integrate metacognition in the everyday discourse of the classroom to foster a metacognitive habit of mind, for instance by examining students' metacognitive evaluation or by using self-questions.

› Consider an undergraduate medical curriculum reform that is aimed at challenging students to develop their higher-order thinking skills and lifelong learning attitude.

## · Reflections on the methodology ·

*"The only people who see the whole picture are the ones who step out of the frame."*
*- S. Rushdie*

Awareness of the overall strengths and limitations of this thesis is essential to value its results. This reflection is additional to the strengths and limitations of each study described in the individual chapters.

The overall methodological rigor is reflected in the use of different methods. Qualitative research methods were used to enrich quantitative research findings and vice versa. Furthermore, most studies were incorporated in authentic educational settings. This allowed us to investigate instructional designs that could be implemented relatively easy in the medical school curriculum if they were found to be effective. A limitation of this approach is that some of our studies were restricted by logistics related to the curriculum, such as time on task during the experiments.

In this thesis, I used critical realism as the philosophical grounding of our studies. I feel comfortably confident when arguing that critical realism is a suitable paradigm for research related to educational neuroscience. Critical realism allowed us to acknowledge the importance of both the natural and the social world (Collier, 1994). This is desirable for educational neuroscience-inspired research as its interdisciplinary nature requires a paradigm that goes beyond the philosophies of education sciences, psychology and neuroscience (Flobakk, 2015). No pre-dominance is given to outcomes from natural sciences or social sciences. For instance, contextual factors were discussed in our scoping review on spaced learning, as well as the potential influence of neuroscientific evidence on the development of new spaced learning formats. In all our studies, we aimed to continuously balance between the relevance of both natural and social factors and how these related to our research design and outcomes.

Critical realism also allowed us to use different methods to investigate different aspects of reality. This is illustrated by this thesis, which includes a variety of qualitative and quantitative methods, i.e. experiments, correlational methods, functional magnetic resonance imaging, semi-structured interviews, thinking aloud, and conventional content analysis. For future researchers, I would like to emphasise the value of using qualitative methods conjoined with quantitative methods, despite the fact that ontologically, the critical realist view on reality is post-positivistic. Some argue that a post-positivistic ontology does not allow for qualitative methods to be used. For example, Gill (2011) states that "one key assumption of qualitative research is that there is not one truth, but multiple truths." This is often linked to the interpretivist epistemology saying that "the job of qualitative researchers… is to acknowledge and report these different realities by relying on the voices and interpretations of the participants through extensive quotes, presenting themes that reflect the words and actions of participants, and advancing evidence of different perspectives on each theme." (Wiltshire, 2018). Contrastingly, critical realists argue that qualitative research may be conducted in order to find the truth, by studying how individual experiences are made manifest by social conditions that are real and existing. Contexts can provide insight into generative mechanisms that may explain certain causalities (Sayer, 1992). As Maxwell (2012) writes, "while critical realism rejects the idea of 'multiple realities', in the sense of independent and incommensurable worlds that are socially constructed by different individuals or societies, it is quite compatible with the idea that there are different valid perspectives on reality." Regarding qualitative methods, this claim permits researchers to move beyond an emphasis on subjective, interpreted experiences and towards uncovering how social relations generate such experiences.

## · Educational neuroscience as a guiding principle ·

*"If the human brain were so simple that we could understand it, we would be so simple that we couldn't." - E. Pugh*

The overall aim of this thesis was to improve health professions education by investigating spaced learning, concept learning and metacognitive learning using an educational neuroscience-inspired approach. Here, I elaborate specifically on the potential of educational neuroscience.

Our studies illustrate that cognitive psychology and neuroscience may guide the development of educational interventions. Examples from cognitive psychology are: the post-decision wagering method, overt learning theory, and metacognitive theory. Additionally, neuroscience informed us on the use of spacing formats. Vice versa, we used educational theories as input for neuroscientific hypotheses to uncover the mechanisms of concept learning. Our fMRI study is an example of how educational neuroscience may also be used to answer fundamental research questions about the science of learning.

Overall, I argue that the value of educational neuroscience-inspired research lies in the interdisciplinary nature of the field and the open-minded view on collaborating with other disciplines. Furthermore, educational neuroscience has an urge to explain underlying mechanisms of behaviour and the field embraces the importance of flexibility in methods. I believe it is this combination of qualities that makes educational neuroscience a valuable addition to currently existing approaches in the health professions education research community.

### Challenges

A researcher may face various challenges that should be taken into account when conducting educational neuroscience-inspired research. First, one may encounter criticism regarding the relevance of neuroscience for education. Some researchers are in favor of a pure psychology approach in order to understand the effect of education on (learning) behaviour (Bowers, 2016; Dougherty & Robey, 2018). However, psychological theories infer hidden causal mechanisms to explain behaviour. These hidden causalities can only be uncovered by studying the underlying brain mechanisms (Thomas, 2018). It is important to emphasise that there is value in going beyond behaviour and moving towards an understanding of the underlying mechanism to improve education, even when behavioural effects are already known (Thomas et al., 2019b):

*"It was known 300 years ago that chewing the bark of the cinchona tree was effective in alleviating the symptoms of malaria. Via the extended contributions of the natural sciences, the U.S. Centers for Disease Control and Prevention now list a range of medicinal treatments for malaria. Understanding mechanisms can improve something that already works."*

Second, it can be difficult not to overinterpret or misinterpret the research findings, leading to conclusions that are too strong (de Bruin, 2016). For instance, it should be considered inappropriate to derive direct implications for teaching practices based on a fMRI study. Instead, translational research is needed on various levels such as the classroom and the curriculum, before one can provide adequate and evidence-informed advice about reforming educational practice. This includes taking into account contextual and political factors. So, keep in mind that conducting educational neuroscience-inspired research may not affect educational practice at first, which does not undermine the quality of the research itself (Gabrieli, 2016):

*"Just as we do not value social neuroscience based on its impact on societal happiness, we should not value educational neuroscience on its ability to reform education."*

Third, it is important to be aware of the seductive allure of educational neuroscience. Various studies have shown that explanations of psychological phenomena are more satisfying to people when they contain neuroscientific information, even when this information is irrelevant (Weisberg et al., 2008). Many underlying reasons have been proposed for this effect. For example, neuroscience may be associated with an apparent objectivity of measurements that are interpreted as more powerful evidence opposed to behavioural research. Or, neuroscience may depict a mind-brain connection that people implicitly believe not to exist (Weisberg et al., 2015). Aside from the possible reasons explaining this effect, the point is that a researcher should be aware that educational neuroscience is an approach, not a goal in itself to satisfy one's environment. The quality of research is not determined by the amount of neuroscience or cognitive psychology that is put into it. There are many examples of interdisciplinary research within different educational domains for instance, that are just as valuable: see also our studies on the multitier approach and refutation texts.

**Opportunities**

Besides the challenges, I will highlight several opportunities that I see for future educational neuroscience-inspired approaches in the health professions education context.

*For the researcher*

In this thesis, we focused our research on spaced learning, concept learning, and metacognitive learning respectively. I would encourage researchers to also consider other topics relevant to health professions education and that may benefit from an educational neuroscience approach. For instance, neuroscientific and psychological knowledge on reward and reinforcement may be helpful in educational research on student motivation and goal orientation. Other interesting topics to explore include cognitive load, emotion, stress, and visualisation (Friedlander et al., 2011).

*For the educator*

Educational neuroscience may help educators to understand and support students' learning behaviours. More firmly, it has been stated that "when teacher training does not include a scientific understanding of learning, this understanding and their practice suffers." (Howard-Jones et al., 2016). Interestingly, beneficial learning effects have been illustrated for learners that were taught about neuroplasticity67. So, educational neuroscience can support learning directly by teaching about the learning brain, or indirectly by encouraging teachers to design interventions in the classroom based on neuroscientific and psychological evidence (Thomas et al., 2019b).

*For the policymaker*

Worldwide, policymakers are enthusiastic when it comes to making evidence-informed decisions about education based on neuroscientific and psychological knowledge about learning (Ansari et al., 2012; Tommerdahl, 2010). Neuroscience is sometimes even perceived as “a remedy for reversing falling standards in education by crafting an evidence base for education and a new learning of science.” (Carew & Magsamen, 2010). Hardiman (2012) claims that educational neuroscience may shift the focus of policymakers from the product of learning to the process of learning, since educational neuroscience is sharply focused on how learners learn. However, policymakers should be aware that this process of learning is highly complex and encompasses neurobiological, cognitive, societal, pedagogical, interactional en ethical aspects. Therefore, close collaborations with researchers and educators are desired. The researchers’ and educators’ communities have expertise in educational and social thinking which contributes to shaping educational goals.

# Appendix

Nederlandse samenvatting

Supplementary

References

List of scientific contributions

Dankwoord

Curriculum Vitae

# Nederlandse samenvatting

**Doel van dit proefschrift**

De kwaliteit van het medisch onderwijs - en daarmee indirect de kwaliteit van zorg - proberen we continu te verbeteren. Daarom streven we al decennia lang naar het ontwikkelen van lesmethodes die wetenschappelijk bewezen effectief zijn. Onderzoekers hebben voorgesteld hierbij te focussen op het leergedrag van onze studenten, zoals benoemd in Hoofdstuk 1. Wat leren ze en hoe leren ze?

Leren beïnvloedt ons brein en ons brein beïnvloedt leren. Ondanks deze vanzelfsprekende relatie tussen het brein en leren, zien we relatief weinig samenwerking tussen de neurowetenschappen, de cognitieve psychologie en de onderwijswetenschappen. Met dit proefschrift slaan we de brug tussen deze wetenschappen om de kwaliteit van het medisch onderwijs naar een hoger niveau te tillen. We onderzoeken 'wat' en 'hoe' studenten leren door de kennis uit de verschillende disciplines te bundelen. Ons doel is om de grenzen tussen de verschillende wetenschappen te vervagen, zodat we effectief bewezen onderwijs, betere zorgprofessionals en uiteindelijk betere zorg kunnen bewerkstelligen.

In dit proefschrift doen we onderzoek naar drie soorten leerprocessen die relevant zijn binnen het medisch onderwijs: *spaced learning, concept learning* en *metacognitive learning*.

**Spaced learning**

Bij *spaced learning* worden leersessies verspreid over de tijd en herhaald, waardoor de leerstof beter beklijft. Ondanks dat neurowetenschappers en psychologen deze manier van leren steunen, wordt *spaced learning* nog maar weinig toegepast in de medische onderwijspraktijk. Gezien medische studenten het moeilijk vinden om de geleerde kennis in hun opleiding te onthouden, zou *spaced learning* voor hen een oplossing kunnen bieden.

Hoofdstuk 2 geeft een literatuuroverzicht van studies die *spaced learning* hebben toegepast in het medisch onderwijs. De uitkomsten van ons onderzoek laten zien dat *spaced learning* vele uiteenlopende vormen kent. Maar, deze vormen zijn veelal niet gestoeld op empirisch of wetenschappelijk bewijs. Daarnaast zijn de toepassingen niet altijd even gedetailleerd beschreven, wat het lastig maakt voor docenten en onderzoekers om de toepassing te repliceren. Tenslotte zien we dat *spaced learning* nu vooral in *online* onderwijsapplicaties wordt gebruikt. Naar aanleiding van deze bevinding stellen wij voor om kansen voor *spaced learning* te benutten in andere onderwijsvormen, bijvoorbeeld *offline* onderwijs dat op locatie wordt gegeven. Daarbij raden we onderzoekers aan om zich te houden aan een meer uniforme definitie van *spaced learning* en om wetenschappelijke literatuur te gebruiken ter ondersteuning. Dit alles zal helpen om *spaced learning* optimaal te integreren in de medische onderwijspraktijk.

Hoofdstuk 3 beschrijft een experiment waarin we *spaced learning* toepassen in klassikaal onderwijs. We hebben een traditioneel hoorcollege, waarin een docent 45 minuten lang vertelt over een onderwerp, omgevormd tot een *spaced* hoorcollege. In deze variant wordt de leerstof teruggebracht tot de essentie. Dit verkorte lesblok wordt door de docent een aantal keer herhaald en tussen de herhalingen zit telkens een pauze van 10 minuten. De uitkomsten laten zien dat de docent en studenten overwegend enthousiast zijn over *spaced learning*. Ondanks dit enthousiasme, leidt *spaced learning* niet tot een betere score van studenten op een geheugentest over de desbetreffende leerstof. Hieruit concluderen wij dat er meer onderzoek nodig is naar de optimale lengte van leersessies en pauzes om het onthouden van de leerstof te bevorderen.

**Concept learning**

*Concept learning* gaat over het leren begrijpen van doorgaans ingewikkelde, abstracte medische concepten zoals bloeddruk en vaatweerstand. Onderzoek laat zien dat het conceptueel begrip onder medische studenten laag is en dat dit lastig te verbeteren is via de huidige lesmethoden. Omdat er vanuit de neurowetenschappen en onderwijswetenschappen verschillende opvattingen zijn over het aanleren van conceptueel begrip, is er nog geen eenduidig antwoord op de vraag hoe we onze medische studenten dit begrip het beste kunnen bijbrengen.

Hoofdstuk 4 beschrijft hoe het gesteld is met het conceptueel begrip van medische studenten. We hebben in dit onderzoek de schriftelijke antwoorden van studenten op medische vraagstukken over bloeddruk, bloedstroom en vaatweerstand geanalyseerd. We zijn in hun antwoorden op zoek gegaan naar misvattingen en misconcepties over deze begrippen. Misvattingen zijn denkfouten die relatief makkelijk zijn op te lossen door studenten zich ervan bewust te maken, denk bijvoorbeeld aan een rekenfout. Misconcepties zijn incorrecte interpretaties die niet stroken met de wetenschappelijke visie. Het zijn juist deze misconcepties die lastig te verhelpen zijn met de huidige onderwijsmethoden. Met dit onderzoek geven we inzicht in de misvattingen en misconcepties die bestaan onder de medische studenten, om zo docenten inzicht te geven in de obstakels die studenten ondervinden bij concept leraning leren en het ontwikkelen van conceptueel begrip.

Hoofdstuk 5 beschrijft de bevindingen van onze studie naar de verschillende opvattingen over *concept learning.* Waar onderwijswetenschappers zeggen dat een misconceptie door te leren uiteindelijk vervangen kan worden door het goede concept, beweren neurowetenschappers dat een misconceptie altijd in de hersenen zal blijven bestaan, ook als je uiteindelijk de juiste conceptie hebt geleerd. Om er in dit laatste geval voor te zorgen dat de misconceptie niet tot uiting komt, beweren neurowetenschappers dat er mechanismen in de hersenen actief zijn die de misconceptie kunnen "inhiberen", oftewel remmen, waardoor alleen de goede conceptie tot uiting komt. Wij onderzoeken deze bewering door de hersenen van medische studenten met en zonder misconcepties te bestuderen met een MRI scanner. De uitkomsten van ons onderzoek laten vooralsnog geen aanwijzingen zien dat remmende mechanismen, zoals beweerd door de neurowetenschappers, een rol spelen bij misconcepties en *concept learning*.

Hoofdstuk 6 beschrijft een experiment waarin we medische studenten onderwerpen aan een lesmethode die concept learning kan stimuleren. Deze methode wordt *peer instruction* genoemd, waarbij studenten in tweetallen een conceptuele vraag bediscussiëren en samen tot een antwoord komen, voordat de vraag klassikaal besproken wordt. We vergelijken deze lesmethode met *self-explanation* waarin studenten individueel nadenken over het antwoord op de vraag en niet onderling mogen discussiëren. De uitkomsten laten zien dat *peer instruction* inderdaad resulteert in meer goede antwoorden dan *self-explanation*. Interessant is dat zelfs twee studenten die van te voren allebei zelfstandig de vraag fout beantwoorden, na *peer instruction* toch vaak op het goede antwoord uitkomen.

**Metacognitive learning**

*Metacognitive learning* gaat over het vergaren van kennis over je eigen denk- en leerproces. Een voorbeeld is het nadenken over welke leerstrategie je het beste in kunt zetten in verschillende situaties. Een ander voorbeeld is het herkennen welk type toets-vraag jou goed ligt en welk type minder en hoe je daarmee omgaat.

Hoofdstuk 7 beschrijft methodes waarmee we bij studenten meten of ze inzicht hebben in hun eigen kennis. Kunnen ze inschatten wat ze weten, maar ook wat ze niet weten? De studenten maken een test waarbij ze bij elke vraag aangeven hoe zeker ze zijn van hun antwoord. Een van de groepen doet dat door een score van 1 tot 5 aan te kruisen op een schaal (waarbij 5 betekent dat een student bijna 100% zeker is dat zijn/haar antwoord goed is). De andere groep doet dit door fictieve punten in te zetten op hun antwoord. Dit laatste principe zou volgens de psychologie een betere maat zijn voor zelfkennis dan het duiden van zekerheid via een score van 1 tot 5. Ondanks deze bewering van de psychologen laat onze uitkomst zien dat de schaal in de onderwijscontext wel degelijk betrouwbaarder lijkt voor het meten van zelfkennis.

Hoofdstuk 8 beschrijft een methode om het niveau van metacognitie bij studenten te diagnosticeren. Er wordt gebruik gemaakt van een multi-tier format. Hierbij wordt er expliciet aan studenten gevraagd naar een antwoord, een uitleg en een inschatting of hun gegeven antwoord en uitleg juist zijn. Voor het rapporteren van deze inschatting is de 5-punts schaal gebruikt die eerder in hoofdstuk 7 is getoetst. Bij onze biomedische studenten zien wij dat er slechts minimale overeenkomst is tussen de inschatting van de studenten wat zij denken te weten en hoe goed ze daadwerkelijk scoren op de vragen. Ze hebben dus weinig inzicht in hun daadwerkelijke begripsniveau.

Hoofdstuk 9 beschrijft een experiment waarin we onderzoeken of een lesmethode, genaamd *refutation texts*, het conceptueel begrip van studenten kan bevorderen door in te spelen op hun zelfkennis. Studenten lezen teksten waarin hun mogelijke misconcepties expliciet worden genoemd en verworpen. Aansluitend wordt een correcte uitleg van het concept gegeven. De uitkomsten laten zien dat door het lezen van *refutation texts* de scores van de studenten en hun zelfkennis verbeteren. Echter, het expliciet benoemen en verwerpen van de misconcepties lijkt geen toegevoegde waarde te hebben.

Hoofdstuk 10 beschrijft een studie waarin we metacognitieve vaardigheden, zoals plannen en evalueren, in kaart brengen. Hiervoor laten we medische studenten hardop denken tijdens het oplossen van medische vraagstukken. We zien een grote verscheidenheid in het gebruik van metacognitieve vaardigheden onder de studenten. In het algemeen besteden de studenten weinig aandacht aan het plannen voorafgaand aan het vraagstuk en aan het evalueren na afloop. Daarnaast zien we dat studenten weinig gebruik maken van effectieve aanwijzingen, zogenoemde cues, die kunnen helpen bij het inschatten van hun zelfkennis. Bij het bevragen van studenten over het belang van zelfinzicht in het medisch onderwijs geeft een deel aan dit als onnodig te beschouwen gezien geneeskundige kennis voor hen vooral draait om onthouden en minder om begrijpen. Een ander deel geeft aan juist meer zelfinzicht te willen ontwikkelen over hun kennisniveau en competentieniveau.

**Afsluiting**

De studies in dit proefschrift richten zich op leerprocessen en instructiemethoden in het medisch onderwijs. In Hoofdstuk 11 worden de implicaties van de resultaten uit dit proefschrift voor de onderwijspraktijk en toekomstig onderzoek in meer detail besproken. We doen onder andere aanbevelingen voor het gebruik van bepaalde instructiemethoden en bediscussiëren het belang van een metacognitieve *mindset* in ons onderwijs. Ten slotte laat dit proefschrift zien dat het samenbrengen van verschillende wetenschapsvelden kan leiden tot innovatief onderwijskundig onderzoek. Ook creëert de discipline overstijgende aanpak tal van mogelijkheden voor onderwijsvernieuwing in het medisch onderwijs. Daarbij vormen de studiemethodes en bevindingen uit dit proefschrift een basis voor verder onderzoek naar leren en instructie.

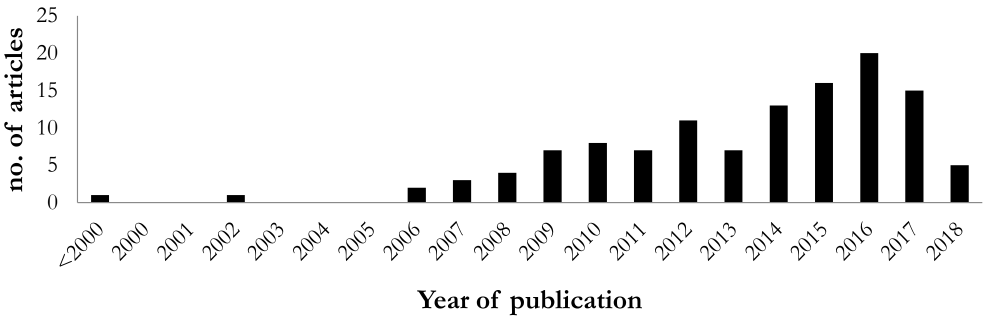
# Supplementary

**Supplementary A** | The used literature search term combinations.

| Spaced [ti] | | Learning [ti] | | Spaced learning [tw] | | Health professions education [tw] |
|---|---|---|---|---|---|---|
| Spaced | AND | Learning | OR | Spaced training | AND | Education, Professional [Mesh] |
| Spacing | | Memory | | Spacing training | | Medical educat* |
| Retriev* | | Training | | Spaced learning | | Professional educat* |
| Repetition* | | Educat* | | Spacing learning | | Teach* |
| Recall | | | | Spaced interval training | | Medical student* |
| Mental recall [Mesh] | | | | Spaced interval learning | | Biomedical student* |
| | | | | Memory training | | Pharmacy student* |
| | | | | Spaced retrieval | | Biology student* |
| | | | | Spaced retrieving | | Dental student* |
| | | | | Retrieval practic* | | Nursing student* |
| | | | | Mass learning | | Undergraduate* |
| | | | | Massed learning | | Graduat* |
| | | | | Spacing effect | | Intern* |
| | | | | Distributed practic* | | University |
| | | | | Spacing efficacy | | Universities |
| | | | | Spacing practice sessions | | Resident* |
| | | | | Spacing of learning | | Health professional* |
| | | | | Spacing repetition | | |
| | | | | Spacing and repetition | | |
| | | | | Spaced educat* | | |
| | | | | Spaced test* | | |
| | | | | Test-enhaned learning | | |
| | | | | Testing effect* | | |

*Within each column all terms are searched with the Boolean operator "OR".*

**Supplementary B |** Frequency distribution of research articles on spaced learning included in this scoping review (*N* = 120). Articles are sorted by year.



**Supplementary C** | Characteristics of included research articles. Except for study location, characteristics were described for original research articles only ($n$ = 103).

| Characteristic | No. (%) |
|---|---|
| **Study location** | |
| United States | 76 (63) |
| Europe | 25 (20) |
| Canada | 8 (7) |
| Australia | 7 (6) |
| Asia | 2 (2) |
| South America | 2 (2) |
| **Study population*** | |
| Medical students | 38 (32) |
| Residents | 27 (23) |
| Other healthcare majors<br>*pharmacy, dental, biomedical sciences, health sciences and prehealthcare* | 18 (15) |
| Other healthcare professions<br>*fellows, physician assistants, nurses, flight paramedics, family doctors, clinicians, program faculty and program directors* | 18 (15) |
| Physicians | 9 (8) |
| **Educational content** | |
| Knowledge<br>*factual, conceptual, procedural* | 74 (72) |
| Skills | 27 (26) |
| N.A. | 4 (4) |
| **Domain** | |
| Clinical | 66 (64) |
| Basic sciences | 9 (9) |
| Miscellaneous | 28 (27) |
| **Subject** | |
| Medical specialties | 60 (58) |
| Guidelines, principles, quality control | 11 (11) |
| Anatomy | 7 (7) |
| Physiology | 5 (5) |
| Molecular biology | 3 (3) |
| Dental education | 2 (2) |
| Miscellaneous<br>*Teaching, Statistics, Nutrition, Resuscitation, Pain-assessment, Physical examination, Sex, Cognitive Behaviour Therapy* | 15 (15) |

**Some studies included multiple study populations.*

**Supplementary D** | Extended overview of spaced learning terms.

| **Terms used without definitions** |
|---|
| Spaced testing |
| Spaced instruction |
| Spaced training |
| Dispersed learning |
| Distributed learning |
| Spaced retrieval practice |
| Spaced studying |
| Repeated retrieval practice |

| Terms with single definition | Definitions |
|---|---|
| Spaced distribution | Educational encounters that are repeated over spaced time intervals. |
| Repeated practice | Practice is distributed over several sessions rather than massed as a single block of training. |
| Spaced approach | Distribution of a fixed amount of teaching hours over a longer time period. |
| Spaced repetition | A learning approach that focuses on reviewing content multiple times over optimized time intervals. |
| Automated spaced repetition | The observation that it is easier to remember information when it is studied multiple times over a long time span, rather than studied once or a few times in a short time span. |
| Structured spaced training | Spaced training; in which training is undertaken with breaks. |
| Interleaved practice | to structure the acquisition of multiple learning sets such that learning alternates between sets. |
| Distributed training | A practice schedule in which periods of training are interspersed with rest periods. |
| Distributed method of learning | instead of learning a French word by continuously reading it off a flashcard, one could memorize it over days or weeks from different sources, for example, by reading newspapers, perusing restaurant menus, and hearing it in conversation. |
| Space repetition learning | Providing small modules of information to learners over time to overcome the normal forgetting that occurs. |

| | |
|---|---|
| Repeated testing | People learn and retain information better through repeated exposure. Actively retrieving content during a test strengthens retention even more. |
| Distributed study | Break the material up into smaller portions that are studied over a period of days. A distributed study strategy is more effective when it includes repetition, with topics being re-visited, so as to reinforce learning. |
| Interactive spaced education | Interactive spaced education (ISE) is an online education system which harnesses the spacing effect, the psychological finding that educational material which is repeatedly presented over spaced intervals is learned and retained more efficiently. |
| Spaced training | Spaced training involves distributing learning tasks and skill acquisition practice sessions over a specified time interval, whereas massed practice occurs over the course of 1 or 2 intense and content-heavy sessions. |
| Interval learning | Learning over time. Seeing the presentation of new materials over time, in bite-sized chunks, and then seeing them again at a later time, particularly as a test. |
| Interval training | Learners practice multiple times for shorter periods in interval training. |

| **Terms with plural definitions** | **Definitions** |
|---|---|
| #1 Spaced practice | Distributing practice over time |
| #2 Spaced practice | Creating a study schedule that spreads study activities out over time. The same amount of repeated studying of the same information spaced out over time will lead to greater retention of that information in the long run, compared with repeated studying of the same information for the same amount of time in one study session. |
| #1 Spaced learning | delivery of brief morsels of information repeated over time intervals, in contrast to massed learning, a bolus of information delivered all at once without breaks |
| #2 Spaced learning | Spaced learning is based on the temporal pattern of stimuli for creating long-term memories. It consists of blocks with highly condensed content that is repeated three times and interrupted by 10–20 min breaks during which distractor activities such as physical activity are performed. |

| | |
|---|---|
| #3 Spaced learning | Learning encounters that are ''spaced'' and ''repeated over time''. Spaced learning differs significantly from other pedagogies because it ''pushes'' short clinical casebased scenarios that take less than 5 minutes to consider to participants' e-mail or hand-held mobile device. |
| #4 Spaced learning | Spaced learning fosters retention effects by a careful adjustment (increasing) of time-intervals between test repetitions. |
| #5 Spaced learning | The insertion of distinct temporal gaps between learning episodes. This is in contrast to massed learning, which refers to learning with little or no time between learning sessions. |
| #1 Retrieval practice | Retrieval practice involves internally recalling previously learned information as opposed to rereading or relistening to information that was previously seen or heard. |
| #2 Retrieval practice | Being tested, pimped, or otherwise asked to exhibit knowledge to which one has been previously exposed. |
| #3 Retrieval practice | In retrieval practice, testing is used as a method of teaching rather than a method of assessment. |
| #4 Retrieval practice | Testing. There are greater gains to be realized with repetition of retrieval practice distributed over longer intervals and interleaved with demands for recall of other concepts or skills. |
| #5 Retrieval practice | "how" one spends time while learning is relevant to the learning science strategy, in contrast to "when" (distributed learning). |
| #6 Retrieval practice | Bringing learned information to mind from long-term memory. |
| #1 Distributed practice | Practice sessions spaced in time. |
| #2 Distributed practice | Practice is distributed over several sessions rather than massed as a single block of training. |
| #3 Distributed practice | Scheduled educational sessions of learning and training interspersed with rest periods. Massed practice refers to a continuous block of training without rest. |
| #4 Distributed practice | Refers to a schedule where periods of practice are interspersed with periods of rest. |
| #5 Distributed practice | Study activities with intervals. |

| | |
|---|---|
| #6 Distributed practice | Distributed practice refers to spacing out ones' practice or relearning materials intermittently over time, whereas massed practice refers to learning materials in one long session that often occurs immediately prior to a learning assessment, e.g. cramming. |
| #7 Distributed practice | Practice interspersed with periods of rest. |
| #8 Distributed practice | Content and understanding of concepts is improved when study is spaced or distributed over a period of time, compared to cramming or massed practice. |
| #9 Distributed practice | A practice regime in which periods of training are interspersed with rest periods. |
| #10 Distributed practice | Knowledge or a skill is taught in multiple sessions separated in time. |
| #11 Distributed practice | Material learnt in brief training workshops decays quickly over time, whereas repetition on many occasions ensures greater retention. |
| #12 Distributed practice | Training in multiple sessions with different intervals. |
| #13 Distributed practice | Several training sessions. |
| #14 Distributed practice | Practice sessions being distributed in time with each session consisting of repeated practice of the exact same procedure. Distribution of practice can refer both to distribution of content into several lessons and to practice sessions being spaced by time, often days or weeks, and is also termed (time) spaced or interval training. |
| #15 Distributed practice | Learning spread over a period of time. Theoretically, teaching in small proportions dispersed over time is better with respect to knowledge and skill retention compared with massed delivery. |
| #16 Distributed practice | Practice sessions are distributed, either as massed or regular teaching sessions. |
| #17 Distributed practice | Involving smaller practice sessions with large interspersed rest periods. |
| #18 Distributed practice | Spacing out of multiple study or practice sessions over an extended period of time. |
| #19 Distributed practice | Creating a study schedule that spreads study activities out over time. The same amount of repeated studying of the same information spaced out over time will lead to greater retention of that information in the long run, compared with repeated studying of the same information for the same amount of time in one study session. |
| #1 Spaced education | The planned repeating of educational encounters over time to enhance knowledge retention. |

| | |
|---|---|
| #2 Spaced education | Spaced education involves spaced repetition of question based educational content, but also includes an adaptive rescheduling algorithm combined with compelling game mechanics. |
| #3 Spaced education | Spaced education is a teaching strategy that delivers educational content electronically over several weeks or months (spaced in time) in a case-based or quiz format. |
| #4 Spaced education | Improved knowledge retention when material is repeated at spaced intervals. |
| #5 Spaced education | The spacing effect is based on the idea of repeating information over spaced intervals of time compared with complete presentation at one time. |
| #6 Spaced education | Online learning that has been demonstrated to increase knowledge retention and impact on behaviour. SE involves participants receiving short multiple-choice questions and feedback via e-mail in a repeating pattern over a number of weeks. |
| #7 Spaced education | Online educational programs that are structured to take advantage of the pedagogical benefits of the spacing effect |
| #8 Spaced education | On-line spaced education programs attempt to improve knowledge retention by harnessing the pedagogical merits of the spacing effect. |
| #9 Spaced education | SPACED education is a novel form of online education based on the 2 core psychology research findings of spacing and testing effects. |
| #10 Spaced education | Online educational programs that are structured to take advantage of the pedagogical benefits of the ‘spacing effect’. |
| #11 Spaced education | Educational programs that are constructed to take advantage of the pedagogical merits of the spacing effect |
| #12 Spaced education | The psychologic finding that training or educational material that is spaced and repeated over time (spaced distribution) results in greater knowledge acquisition and retention of content than when the same material is presented at a single time point (mass distribution). |
| #13 Spaced education | Educational content delivered in small quantities and repeated over time, through electronic distribution |
| #14 Spaced education | A conceptual educational framework that relies on the delivery of succinct pieces of information related to a specific content area spaced out over time. |

| | |
|---|---|
| #15 Spaced education | Involves participants' receiving short multiple-choice questions and feedback via e-mail in a repeating pattern over a number of weeks. |
| #16 Spaced education | Online educational programs that are structured to take advantage of the pedagogical benefits of the spacing effect, in which periodically repeated, educational encounters lead to improved knowledge attainment and retention compared with a single "bolus" educational opportunity. |
| #17 Spaced education | An elearning platform in which "information" is presented and repeated over spaced intervals and is learned and retained more effectively. |
| #18 Spaced education | Educational encounters are spaced and repeated over a defined period. |
| #19 Spaced education | A web-based, student-directed learning tool. |

**Supplementary E** | Retention-test test analyses.

| | Mean | SD |
|---|---|---|
| Total score (points) | 9.55 | 3.82 |
| Total score (%) | 32.94 | 13.18 |
| KR-20 alpha | 0.74 | |
| Question nr. | p-value | Rir |
| Question 1 | 0.00 | 0.00 |
| Question 2 | 0.11 | 0.46 |
| Question 3 | 0.65 | 0.20 |
| Question 4 | NA | NA |
| Question 5 | 0.88 | 0.33 |
| Question 6 | 0.88 | 0.17 |
| Question 7 | 0.61 | 0.27 |
| Question 8 | 0.34 | 0.38 |
| Question 9 | 0.04 | 0.16 |
| Question 10 | 0.07 | 0.07 |
| Question 11 | 0.05 | 0.36 |
| Question 12 | 0.88 | 0.22 |
| Question 13 | 0.28 | 0.36 |
| Question 14 | 0.03 | 0.27 |
| Question 15 | 0.06 | 0.11 |
| Question 16 | 0.36 | 0.25 |
| Question 17 | 0.19 | 0.25 |
| Question 18 | 0.66 | 0.27 |
| Question 19 | 0.52 | 0.39 |
| Question 20 | 0.28 | 0.28 |
| Question 21 | 0.11 | 0.40 |
| Question 22 | 0.05 | 0.32 |
| Question 23 | 0.89 | 0.34 |
| Question 24 | 0.14 | 0.24 |
| Question 25 | 0.36 | 0.27 |
| Question 26 | 0.43 | 0.33 |
| Question 27 | 0.22 | 0.07 |
| Question 28 | 0.04 | -0.09 |
| Question 29 | 0.50 | 0.30 |
| Question 30 | 0.25 | 0.23 |

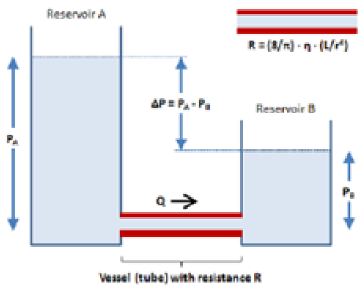
**Supplementary F** | Information sheet included in the study protocol (Dutch).

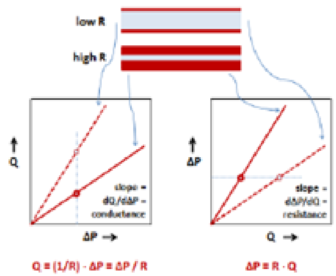
**BEGRIPPEN en FORMULES**

**De figuren op dit blad illustreren de begrippen weerstand en compliantie, je krijgt nu een aantal vragen over deze begrippen. Je mag dit blad bij het beantwoorden van de vragen blijven raadplegen. Gebruik de achterkant als notitieblad. Je hoeft dit blad niet in te leveren. Neem nu eerst 2 minuten om dit blad nogmaals te bestuderen.**

Weerstand karakteriseert de relatie tussen drukval en stroming door een vat (buis).
Drukval is het verschil in druk tussen begin en eind van het vat.

R weerstand
ΔP drukval = driving pressure
Q stroming = flow



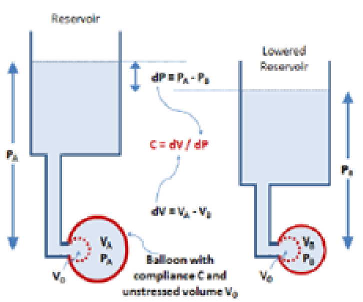



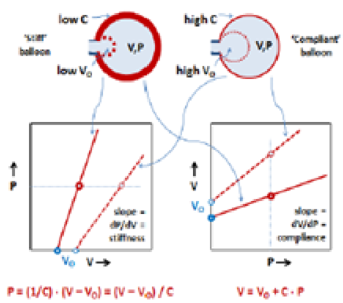

In de medische fysiologie wordt voor druk meestal de eenheid mmHg gebruikt en voor stroming ml/s.
De eenheid voor weerstand is dan dus mmHg/(ml/s).
Maar soms worden ook andere eenheden gebruikt, zoals $cmH_2O$ voor druk en L/min voor stroming. In dat geval wordt de eenheid voor weerstand $cmH_2O/(L/min)$
NB: weerstand = 1/geleiding ofwel resistance = 1/conductance

---

Compliantie karakteriseert, samen met het unstressed volume, de relatie tussen druk (transmurale druk) en volume van een compartiment (ballon).
Transmurale druk is het verschil in druk tussen binnen en buiten het compartiment.
'Unstressed volume' is het volume waarmee het compartiment gevuld kan worden, zonder dat de wand opgerekt wordt en dus zonder dat de druk gaat stijgen (volume bij druk = 0).

C compliantie
$V_0$ unstressed volume
P druk
V volume




In de medische fysiologie wordt voor druk meestal de eenheid mmHg gebruikt en voor volume (en voor unstressed volume) ml. De eenheid voor compliantie is dus ml/mmHg.
NB: compliantie = 1/stijfheid = 1/elastantie (dus lage compliantie = hoge stijfheid)

**Supplementary G** | Clusters significantly more activated in understanders and misunderstanders.

| Group | Condition | Region | MNI coordinates (x; y; z, mm) | Cluster size (voxels) | Cluster P |
|---|---|---|---|---|---|
| Understanders | Scientific conception | R middle temporal area (BA21) | 62;-34;-14 | 26334 | .0004 |
| | | R supplementary motor area (BA6) | 60;-4;46 | | |
| | | R anterior cingulate cortex (BA32) | 6;-4;46 | | |
| | | L anterior cingulate cortex (BA24) | -6;-8;46 | | |
| | | R superior parietal lobe/precuneus (BA7) | 26;-56;45 | | |
| | | R putamen (BA49) | 38;-2;-2 | 3461 | .0204 |
| | | R supplementary motor area (BA6) | 64;8;16 | | |
| | | R inferior frontal gyrus (BA44) | 62;12;14 | | |
| | | R inferior temporal gyrus (BA37) | 38;-50;-22 | 2502 | .0338 |
| | Misconception | L anterior cingulate cortex (BA32) | -8;14;34 | 27720 | .0002 |
| | | R superior parietal lobe/precuneus (BA7) | 8;-76;50 | | |
| | | L superior parietal lobe/precuneus (BA7) | -20;-70;50 | | |
| | | R ventrolateral prefrontal cortex (BA47) | 34;30;-22 | 3349 | .0172 |
| | | R inferior frontal gyrus (BA44) | 52;8;-2 | | |
| | | R anterior prefrontal cortex (BA10) | 28;54;-18 | | |
| | Positive control | R inferior temporal gyrus (BA37) | 32;-40;-24 | 32120 | .0002 |
| | | R anterior cingulate cortex (BA32) | 8;8;34 | | |
| | | R inferior frontal gyrus (BA44) | 62;12;14 | | |
| | | L inferior parietal lobe (BA40) | -38;-38;48 | | |
| | | L transverse temporal area (BA41) | -40;-26;10 | | |
| | | R thalamus (BA50) | 16;-16;8 | | |

| | | | | | |
|---|---|---|---|---|---|
| | Negative control | R inferior temporal gyrus (BA37) | 38;-42;-24 | 31446 | .0002 |
| | | R anterior cingulate cortex (BA24) | -8;0;42 | | |
| | | L inferior parietal lobe (BA40) | -44;-32;12 | | |
| | | R inferior frontal gyrus (BA44) | 60;14;10 | | |
| | | L transverse temporal area (BA41) | -48;-20;10 | | |
| | | L anterior prefrontal cortex (BA10) | -23;59;-14 | | |
| | | R ventrolateral prefrontal cortex (BA47) | -26;18;-20 | | |
| Misunderstanders | Scientific conception | R supplementary motor area (BA6) | 62;8;4 | 65211 | .0002 |
| | | L supplementary motor area (BA6) | -54;-4;40 | | |
| | | R inferior temporal gyrus (BA37) | 52;-54;6 | | |
| | | R anterior cingulate cortex (BA24) | 8;0;42 | | |
| | | L dorsolateral prefrontal cortex (BA9) | -38;37;29 | | |
| | | R inferior frontal gyrus (BA44) | 60;14;10 | | |
| | Misconception | R inferior temporal gyrus (BA37) | 26;-32;26 | 74691 | .0002 |
| | | R angular gyrus (BA39) | 40;-64;42 | | |
| | | R visual association area (BA18) | 2;-80;-6 | | |
| | | R superior temporal gyrus (BA22) | 54;2;-6 | | |
| | | R anterior cingulate cortex (BA24) | 8;0;42 | | |
| | | R dorsolateral prefrontal cortex (BA9) | 34;38;29 | | |
| | | R inferior frontal gyrus (BA44) | 60;14;10 | | |
| | Positive control | R inferior temporal gyrus (BA37) | 30;-42;-24 | 67524 | .0002 |
| | | R transverse temporal area (BA41) | 40;-32;10 | | |
| | | L superior parietal lobe/precuneus (BA7) | -24;-46;52 | | |
| | | R superior temporal gyrus (BA22) | 64;-42;16 | | |
| | | R anterior cingulate cortex (BA32) | 8;3;42 | | |
| | | L dorsolateral prefrontal cortex (BA9) | -38;37;29 | | |
| | | R inferior frontal gyrus (BA44) | 60;14;10 | | |

| | | | | | |
|---|---|---|---|---|---|
| | Negative control | R inferior temporal gyrus (BA37) | 38;-42;-24 | 65573 | .0002 |
| | | R anterior cingulate cortex (BA32) | 8;3;42 | | |
| | | R inferior parietal lobe (BA40) | 50;-28;32 | | |
| | | R medial prefrontal cortex (BA8) | 12;20;32 | | |
| | | L anterior prefrontal cortex (BA10) | -28;50;-11 | | |
| | | L angular gyrus (BA39) | -28;-60;34 | | |
| | | R thalamus (BA50) | 16;-14;2 | | |
| | | L dorsolateral prefrontal cortex (BA9) | -38;37;29 | | |
| | | R inferior frontal gyrus (BA44) | 60;14;10 | | |
| Misunderstanders > Understanders | Scientific conception | L putamen (BA49) | -24; -8; -28 | 1163 | .0024 |
| | Misconception | L putamen (BA49) | -34;-6;-10 | 703 | .0068 |

*Cluster-defining threshold p < .001, cluster extent threshold p < .05, FWE corrected (random effect analysis). Z-max for all clusters: 3.54. Coordinates are based on the most significant voxel in the cluster.*

**Supplementary H** | Clusters significantly more activated in a certain comparison of conditions.

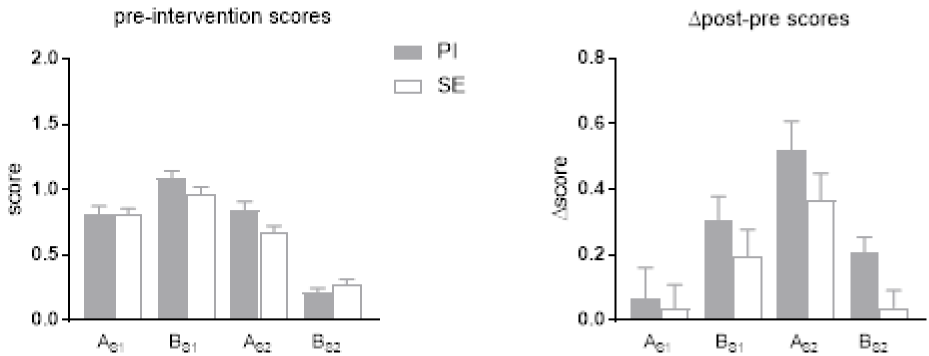
| Group | Condition | Region | MNI coordinates (x; y; z, mm) | Cluster size (voxels) | Cluster P |
|---|---|---|---|---|---|
| Understanders | Scientific conception > Positive control | L inferior parietal lobe (BA40) | -38;-40;38 | 4088 | .0006 |
| | | R superior parietal lobe/precuneus (BA7) | 26;-64;42 | | |
| | | L superior parietal lobe/precuneus (BA7) | -18;-70;52 | | |
| | Scientific conception > Negative control | R visual association area (BA18) | 38;-92;18 | 3768 | .0004 |
| | | R superior parietal lobe/precuneus (BA7) | 26;-58;62 | | |
| | | L superior parietal lobe/precuneus (BA7) | -30;-58;60 | | |
| | | R occipito-temporal cortex (BA19) | 38;-82;18 | | |
| | Misconception > Positive control | L inferior parietal lobe (BA40) | -42;-44;48 | 2738 | .0006 |
| | | L superior parietal lobe/precuneus (BA7) | -20;-72;50 | | |
| | | L postcentral gyrus (BA1) | -50;-34;52 | | |
| | | R superior parietal lobe/precuneus (BA7) | 22;-66;48 | 1849 | .0018 |
| | | R angular gyrus (BA39) | 46;-60;38 | | |
| | | R inferior parietal lobe (BA40) | 44;-52;52 | | |
| | | L orbitofrontal cortex (BA11) | -8;50;-18 | 610 | .0414 |
| | | L ventrolateral prefrontal cortex (BA47) | -50;42;-16 | | |
| | Misconception > Negative control | L superior parietal lobe/precuneus (BA7) | -18;-72;50 | 2717 | .0002 |
| | | R superior parietal lobe/precuneus (BA7) | 18;-70;52 | | |

| | | | | | |
|---|---|---|---|---|---|
| | Negative control > Positive control | R orbitofrontal cortex (BA11) | 20; 38; -16 | 1446 | .0040 |
| | | R ventrolateral prefrontal cortex (BA47) | 36;44;-16 | | |
| | | R anterior prefrontal cortex (BA10) | 38; 56; -14 | | |
| | | L anterior prefrontal cortex (BA10) | -26;56;-14 | | |
| | | R angular gyrus (BA39) | 46;-58;32 | 1147 | .0058 |
| | | R inferior parietal lobe (BA40) | 44;-48;44 | | |
| | | L anterior temporal lobe (BA38) | -42;14;-36 | 764 | .0140 |
| | | L middle temporal gyrus (BA21) | -56;-6;-22 | | |
| | | | | | .0180 |
| | | L angular gyrus (BA39) | -44;-60;46 | 658 | |
| | | L inferior parietal lobe (BA40) | -48;-40;40 | | |
| | | | | | .0274 |
| | | R inferior temporal gyrus (BA37) | 66;-48;-14 | 542 | |
| | | R middle temporal gyrus (BA21) | 64;-40;-10 | | |
| Misunderstanders | Scientific conception > Positive control | L inferior parietal lobe (BA40) | -64; -28; 42 | 5657 | .0006 |
| | | L superior parietal lobe/precuneus (BA7) | -36; -58; 64 | | |
| | | R superior parietal lobe/precuneus (BA7) | 18; -66; 60 | | |
| | | L postcentral gyrus (BA1) | -48; -32; 58 | | |
| | Scientific conception > Negative control | L visual association area (BA18) | -26; -82; 0 | 4101 | .0002 |
| | | L superior parietal lobe/precuneus (BA7) | -22; -66; 58 | | |
| | | R superior parietal lobe/precuneus (BA7) | 22; -68; 58 | | |
| | | R visual association area (BA18) | 34; -84; 6 | 1612 | .0108 |

| | | | | |
|---|---|---|---|---|
| Misconception > Positive control | R inferior temporal gyrus (BA37) | 54; -54; -14 | 14729 | .0002 |
| | L visual association area (BA18) | -24; -86; 0 | | |
| | R visual association area (BA18) | 36; -84; 0 | | |
| | L superior parietal lobe/precuneus (BA7) | -16; -68; 54 | | |
| | L occipito-temporal cortex (BA19) | -44; -78; -14 | | |
| Misconception > Negative control | R visual association area (BA18) | 16; -98; -10 | 10352 | .0002 |
| | R occipito-temporal cortex (BA19) | 40; -84; 24 | | |
| | R superior parietal lobe/precuneus (BA7) | 24; -68; 52 | | |
| | L superior parietal lobe/precuneus (BA7) | 22; -66; 54 | | |
| Negative control > Positive control | L inferior parietal lobe (BA40) | -46; -40; 50 | 1768 | .0022 |
| | L superior parietal lobe/precuneus (BA7) | -30; -62; 38 | | |
| | L angular gyrus (BA39) | -40; -50; 36 | | |
| Misconception > Scientific conception | L angular gyrus (BA39) | -38; -66; 38 | 1103 | .0132 |
| | R visual association area (BA18) | 38; -86; -4 | 1010 | .0174 |
| | L inferior temporal gyrus (BA37) | -66; -40 ;-16 | 876 | .0244 |
| | L middle temporal area (BA21) | -70; -28; -8 | | |

*Cluster-defining threshold $p < .001$, cluster extent threshold $p < .05$, FWE corrected (random effect analysis). Z-max for all clusters: 3.54. Coordinates are based on the most significant voxel in the cluster.*

**Supplementary I** | Pre-test scores on conceptual problems in seminars 1 and 2. Pre-post scores were examined separately for all four questions (two concepts in two seminars) to assess if efficacy of active learning was consistent among these exercises. Average pre scores are shown for PI and SE groups (left panel). Lowest scores were found for exercise $B_{s2}$. The relative change in scores was positive for all exercises in both SE and PI groups, with higher gains for PI in all cases (right panel). Accordingly, a significant increase of 0.17 was found on exercise $B_{s2}$ scores in the second seminar between PI and SE groups ($t(151)=3.206$, $p = .023$). The highest performance gains for single exercises were observed in the PI condition, showing the largest increase of 0.52±0.09 for $A_{s2}$.

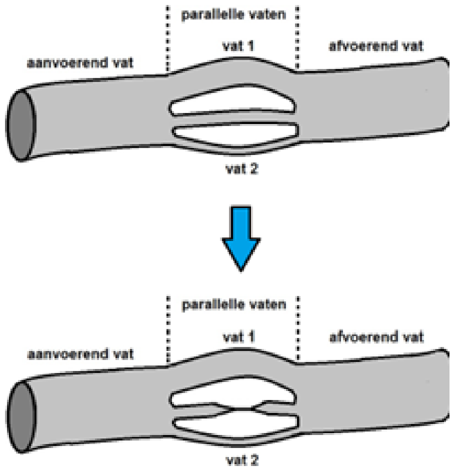


**Supplementary J |** Students' individual scores on conceptual problems before and after intervention. The conceptual problems consisted of two related questions each yielding 1 point, thus a maximal score of 2. This scheme shows the distribution of total scores pre and post intervention for all students participating in the protocol. The total scores (0, 1 or 2 points) are indicated in the circles. The white, grey and dark areas indicate, respectively, students with a pre-test score of 0, 1 or 2 points. The numbers in the bars indicate the number of students with a specific score. Data for both seminars is depicted separately to illustrate the difference in difficulty level between the seminars. Seminar 2 had more students with score 0 and fewer with scores 1 or 2, but similar shifts towards higher scores post intervention.

**Supplementary K** | Multi-tier question (in Dutch) with 3-tiers: Yes/No, Explanation, and Confidence.

**Vraagstuk 2**

Een chirurg ziet in het onderbeen van een patiënt deze vaatstructuur: drie parallelle vaten met een aanvoerend en een afvoerend vat. De flow door vat 1 is het dubbele van de flow door vat 2.




Dan klemt de chirurg het middelste vat af waardoor er geen flow meer door dat vat gaat. De chirurg beweert dat de flow door vat 1 nog steeds het dubbele is van de flow door vat 2. Je mag aannemen dat de diameters van vat 1 en vat 2 niet veranderen.

Heeft de chirurg gelijk?

- ☐ Ja
- ☐ Nee

Wat is de beste verklaring of het belangrijkste argument voor je antwoord?

- ☐ Omdat de flow verdeling bepaald wordt door de verhouding van de weerstanden
- ☐ Omdat de flow via het vat met de laagste weerstand gaat
- ☐ Omdat de totale flow constant blijft
- ☐ Omdat de totale weerstand toeneemt en dus de flow in beide vaten evenveel afneemt

Hoe zeker ben je dat je antwoord inclusief verklaring correct is?

| ☐ | ☐ | ☐ | ☐ | ☐ |
|---|---|---|---|---|
| Heel onzeker (was gok) | Vrij onzeker | Ik twijfel | Vrij zeker | Heel zeker (bijna 100%) |

**Supplementary L** | Refutation text (in Dutch) with a refutation element, correct answer, and explanation.

Sommige mensen zeggen dat de flow de weg van de minste weerstand kiest en dat dus de flow die na het afklemmen niet meer door het middelste vat kan, door vat 1 zal gaan. Dat is niet correct omdat flow niet via de weg van de minste weerstand gaat, maar zich verdeelt afhankelijk van de weerstanden. Het vat met de laagste weerstand krijgt de meeste flow, maar ook een vat met een hogere weerstand krijgt flow.

Het juiste antwoord is:
*Ja, de chirurg heeft gelijk, omdat de flow verdeling tussen vat 1 en vat 2 bepaald wordt door de verhouding van de weerstanden (en die is niet veranderd).*

De drukval is voor alle parallelle vaten hetzelfde. Flow is drukval/weerstand, dus de verhouding tussen de flows wordt bepaald door de verhouding tussen de weerstanden: het vat met de laagste weerstand krijgt de meeste flow. Omdat vooraf de flow in vat 1 het dubbele is van de flow door vat 2 is dus blijkbaar de weerstand van vat 1 de helft van de weerstand van vat 2. Nadat het middelste vat is afgeklemd verandert wellicht de drukval, maar deze is wel nog steeds voor alle parallelle vaten gelijk. De weerstanden van vat 1 en 2 blijven hetzelfde, dus ook hun verhouding. En dus gaat er nog steeds tweemaal zoveel flow door vat 1 als door vat 2.
Wellicht neemt de flow door vat 1 toe, maar dat is dan ook voor vat 2 het geval. Echter de toename in vat 2 is maar de helft van de toename in vat 1, omdat de weerstand van vat 2 tweemaal zo hoog is als van vat 1. Hoewel dus beide flows kunnen veranderen, blijft de verhouding tussen de twee flows hetzelfde.

**Supplementary M** | Interview guide.

**Evaluating the thinking aloud assignments**

What was your experience with doing these assignments?

o How come?

o How did you feel?

o Which parts did you find easy/difficult?

o How did you estimate your confidence after finishing an assignment?

**Metacognitive knowledge**

Where do you start when you get such conceptual assignments?

o What do you do when you do not understand the assignment, or part of the assignment?

o Is the assignment finished the moment you write down your answer?

o Can you tell me what you do afterwards?

Do you use specific strategies while solving this type of assignments?

o Does such a strategy depend on the type of question? How?

o Did you ever change strategies for different types of questions? Why?

Can you tell me what you strenghts and weaknesses are regarding solving this type of conceptual assignments?

**Metacognition in the curriculum**

Can you describe any situations or name ways through which you obtain insight in how well you understand something?

If we would improve the curriculum in such a way that it becomes easier for you to estimate how well you understanding something, does that have added value for you?

o How do you think we should improve the curriculum?

**Supplementary N** | Template thinking aloud and interview.

### Metacognitive experiences

| MAIN THEME | SUB THEME | EXAMPLE |
|---|---|---|
| Feeling of knowing | | 'Eureka.'<br>'I do not think so, but I do not know why.'<br>'Intuitively, or based on my feeling it has to be.. maybe I have the feeling I am making a mistake.' |
| Feeling of familiarity | | 'Yes, I recognise this image. I think I have seen this before.'<br>'I recognise this from year 1.' |
| Feeling of confidence | Sure | 'I am quite sure about this.' |
| | Unsure | 'Yes. No. I am not quite sure.' |
| | Doubting | 'Yes, now I start doubting.' |
| Feeling of difficulty | | 'O god, this is difficult.'<br>'O yes, I was afraid of this [type of exercise] already.' |

### Metacognitive skills

| MAIN THEME | SUB THEME | EXAMPLE |
|---|---|---|
| Planning | Allocating resources | 'I will start by reading the formulas.' |
| | Identifying problem | 'OK, so it is about...' |
| | Creating overview | 'I will write that down for now.. so I have an overview.' |
| Monitoring | Rereading | Rereading a paragraph. |
| | Summarising | 'I will write summarise this because I might need it later.' |

| | |
|---|---|
| Eliminating | ‘So then we have the option left..’<br>‘So this one can be eliminated for sure..’ |
| Selecting new source | Allocates a new resourse, e.g. formula sheet, halfway through the assignment. |
| Using mnemonic | Using a mnemonic to explain the situation in layman language. |
| Using numbers | ‘OK, so for example, we fill in numbers here..’ |
| Using visuals | ‘Let’s have a look. There is an image here.’ |
| Using answer options | ‘OK, I will read the answer options, because that might help me figure it out..’ |
| Switching strategy | ‘I now see that..’ |
| Identifying adequacy of information | ‘ Hm, this formula is not useful at the moment.’ |
| Self-questioning | ‘Is this right what I am saying?’ |
| Goal-checking | ‘What was the problem again?’ |
| Self-explanation | Explaining to him/herself why the answer is incorrect/correct before committing to an answer. |
| Goal-free search | ‘Maybe it has something to do with..’ |
| Goal-directed search | ‘I will search in the formula sheet for resistance, flow and pressure difference.’ |

| | | |
|---|---|---|
| | Judgement of learning | 'These [the answer options] are all not the same compared to my answer. Then I have to look again. Maybe I am wrong.'<br><br>'I will just make a decision based on what sounds logical.. and based on what I remembered, not based on the reasoning and the process behind it'. |
| Evaluation | Checking answer | ' I will just check everything again, because..' |
| | Elaborating on confidence | 'So if the resistance decreases, so.. yes I am very sure about my answer.' |

**Metacognitive knowledge**

| MAIN THEME | SUBTHEME |
|---|---|
| Of self | Formulas/numbers |
| | Visual representation |
| | Speed of reasoning |
| | Not thinking 'straight' |
| | Insecurity |
| | Lack of knowledge/brainpower |
| | Compensating weaknesses |
| | Influence of environment |
| | Easy |
| | Overview of the stiutation |
| | Abstract thinking |
| | Estimating |
| Of Task | Different strategies for figures/formulas |
| | Different strategies for facts/cooncepts |
| Of Strategy | Not aware of strategy |
| | Eliminating options |
| | Using visual representation |
| | Highliting |
| | Thinking out loud |

| | |
|---|---|
| | Identifying the problem |
| | Goal-checking |
| | Checking previous assumptions |
| | Good reading |
| | Using numbers |
| | Creating overview |
| | Filling in formulas |
| | Tackle the problem from different angles |
| | Evaluating |
| Cues for confidence | Feeling |
| | Time |
| | Familiarity/previous experiences |
| | Reasoning |
| | Deduction |
| | Representation (figures/formulas) |
| | Combination of reasoning and feelings |

# References

# A

Ahopelto, I., Mikkila-Erdmann, M., Olkinuora, E., & Kaapa, P. (2011). A follow-up study of medical students' biomedical understanding and clinical reasoning concerning the cardiovascular system. *Advances in Health Sciences Education, 16*(5), 655-668.

Ajay, S. M., & Bhalla, U. S. (2004). A role for ERKII in synaptic pattern selectivity on the time-scale of minutes. *European Journal of Neuroscience, 20*(10), 2671-2680.

Akdemir, A., Zeybek, B., Ergenoglu, A. M., Yeniel, A. O., & Sendag, F. (2014). Effect of spaced training with a box trainer on the acquisition and retention of basic laparoscopic skills. *International Journal of Gynecology & Obstetrics, 127*(3), 309-313.

Akkaraju, S. (2016). The Role of Flipped Learning in Managing the Cognitive Load of a Threshold Concept in Physiology. *Journal of Effective Teaching, 16*(3), 28-43.

Alba, J. W., & Hasher, L. (1983). Is memory schematic? *Psychological Bulletin, 93*(2), 203-231.

Allaire-Duquette, G., Belanger, M., Grabner, R. H., Koschutnig, K., & Masson, S. (2019). Individual differences in science competence among students are associated with ventrolateral prefrontal cortex activation. *Journal of Neuroscience Research, 97*(9), 1163-1178.

Alparslan, C., Tekkaya, C., & Geban, O. (2003). Using the conceptual change instruction to improve learning. *Journal of Biological Education, 37*(3), 133-137.

Al-Rawi, W., Easterling, L., & Edwards, P. C. (2015). Development of a mobile device optimized cross platform- compatible oral pathology and radiology spaced repetition system for dental education. *Journal of Dental Education, 79*(4), 439-447.

Andersen, S. A. W. (2016). Virtual reality simulation training of mastoidectomy: studies on novice performance. *Danish Medical Journal, 63*:B5277.

Andersen, S. A. W., Konge, L., Caye-Thomasen, P., & Sorensen, M. S. (2016a). Retention of mastoidectomy skills after virtual reality simulation training. J*AMA Otolaryngology–Head & Neck Surgery, 142*(7), 635-640.

Andersen, S. A. W., Mikkelsen, P. T., Konge, L., Caye-Thomasen, P., & Sorensen, M. S. (2016b). Cognitive load in distributed and massed practice in virtual reality mastoidectomy simulation. *The Laryngoscope, 126*(2), E74- E79.

Andersen, S. A. W., Konge, L., Caye-Thomasen, P., & Sorensen, M. S. (2015). Learning curves of virtual mastoidectomy in distributed and massed practice. J*AMA Otolaryngology–Head & Neck Surgery, 141*(10), 913-918.

Anderson, J., (1979). For multiple choice questions. *Medical Teacher, 1*(1), 37-42.

Anderson, L. W., Krathwohl, D. R., Airiasian, W., Cruikshank, K. A., Mayer, R. E., & Pintrich, P. R. (2001). A taxonomy for learning, teaching and assessing: A revision of Bloom's Taxonomy of educational outcomes. NY: Longman.

Anderson, G. L., Passmore, J. C., Wead, W. B., Falcone, J. C., Stremel, R. W., & Schuschke, D. A. (2011). Using "active learning" methods to teach physiology. *Medical Science Educator, 21*(1), 8-20.

Anderson, R. C., & Pearson, P. D. (1984). A schema-theoretic view of basic processes in reading comprehension. In: Pearson P. D., Barr R., Kamil M.L., Mosenthal P., editors. Handbook of reading research. (pp. 255-291). NY: Longman.

Ansari, D., Coch, D., & De Smedt, B. (2011). Connecting Education and Cognitive Neuroscience: Where will the journey take us? *Educational Philosophy and Theory, 43*(1), 37-42.

Ansari, D., De Smedt, B. & Grabner, R. H. (2012). Neuroeducation–a critical overview of an emerging field. *Neuroethics 5*(2), 105-117.

Antonoff, M. B., & D'Cunha, J. (2011). Retrieval practice as a means of primary learning: Socrates had the right idea. *Seminars in Thoracic and Cardiovascular Surgery, 23*(2), 89-90.

Archer, J. C. (2010). State of the science in health professional education: effective feedback. *Medical Education, 44*(1), 101-108.

Arksey, H., & O'Malley, L. (2005). Scoping studies: towards a methodological framework. *International Journal Of Social Research Methodology, 8*(1), 19-32.

Artino, Jr A. R., Cleary, T. J., Dong, T., Hemmer, P. A., & Durning, S. J. (2014). Exploring clinical reasoning in novices: a self-regulated learning microanalytic assessment approach. *Medical Education, 48*(3), 280-291.

Artino, Jr A. R., Dong, T., DeZee, K. J., Gilliland, W. R., Waechter, D. M., Cruess, D., & Durning, S. J. (2012). Achievement goal structures and self-regulated learning: relationships and changes in medical school. *Academic Medicine, 87*(10), 1375-1381.

Augustin, M. (2014). How to learn effectively in medical school: test yourself, learn actively, and repeat in intervals. *The Yale Journal of Biology and Medicine, 87*(2), 207-212.

Ayyub, A., & Mahboob, U. (2017). Effectiveness of Test-Enhanced Learning (TEL) in lectures for undergraduate medical students. *Pakistan Journal of Medical Sciences, 33*(6), 1339-1343.

## B

Badenhorst, E., Hartman, N., & Mamede, S. (2016). How biomedical misconceptions may arise and affect medical students' learning: A review of theoretical perspectives and empirical evidence. *Health Professions Education, 2*(1), 10-17.

Badenhorst, E., Mamede, S., Hartman, N., & Schmidt, H. G. (2015). Exploring lecturers' views of first-year health science students' misconceptions in biomedical domains. *Advances in Health Sciences Education, 20*(2), 403-420.

Baghdady, M., Carnahan, H., Lam, E. W., & Woods, N. N. (2014). Test-enhanced learning and its effect on comprehension and diagnostic accuracy. *Medical Education, 48*(2), 181-188.

Barnett, S. M., & Ceci, S. J. (2002). When and where do we apply what we learn? A taxonomy for far transfer. *Psychological Bulletin, 128*(4), 612-637.

Barrington, M. J., Viero, L. P., Kluger, R., Clarke, A. L., Ivanusic, J. J., & Wong, D. M. (2016). Determining the learning curve for acquiring core sonographic skills for ultrasound-guided axillary brachial plexus block. *Regional Anesthesia & Pain Medicine, 41*(6), 667-670.

Barsoumian, A. E., & Yun, H. C. (2018). Augmenting fellow education through spaced multiple-choice questions. *Military Medicine, 183*(1-2), e122-e126.

Bartley, J. E., Riedel, M. C., Salo, T., Boeving, E. R., Bottenhorn, K. L., Bravo, E. I., . . . Sutherland, M. T. (2019). Brain activity links performance in science reasoning with conceptual approach. *NPJ Science of Learning, 4*(1), 1-8.

Begg, I., Duft, S., Lalonde, P., Melnick, R., & Sanvito, J. (1989). Memory predictions are based on ease of processing. *Journal of Memory and Language, 28*(5), 610-632.

Behr, M. J., Harel, G., Post, T. R., & Lesh, R. (1992). Rational number, ratio, and proportion. In: D. A. Grouws (Ed.), Handbook of research on mathematics teaching and learning: A project of the National Council of Teachers of Mathematics (pp. 296–333). NY: Macmillan Publishing Co, Inc.

Bekkink, M. O., Donders, R., van Muijen, G. N., & Ruiter, D. J. (2012). Challenging medical students with an interim assessment: a positive effect on formal examination score in a randomized controlled study. *Advances in Health Sciences Education, 17*(1), 27-37.

Bell P., & Volckmann D. (2011). Knowledge surveys in general chemistry: confidence, overconfidence, and performance. *Journal of Chemical Education, 88*(11), 1469-1476.

Bergman, E., de Feijter, J., Frambach, J., Godefrooij, M., Slootweg, I., Stalmeijer, R., & van der Zwet, J. (2012). AM last page: A guide to research paradigms relevant to medical education. *Academic Medicine, 87*(4), 545.

van Berkel, H., & Bax, A. (2014). Toetsen: Toetssteen of dobbelsteen. In: Toetsen in het hoger onderwijs (pp. 15- 27). Bohn Stafleu van Loghum, Houten.

Berkhout, J. J., Helmich, E., Teunissen, P. W., van den Berg, J. W., van der Vleuten, C. P., Jaarsma, A. D. C. (2015). Exploring the factors influencing clinical students' self-regulated learning. *Medical Education, 49*(6), 589-600.

Berner, E. S., & Graber, M. L. (2008). Overconfidence as a cause of diagnostic error in medicine. *The American Journal of Medicine, 121*(5), S2-S23.

Bjerrum, A. S., Eika, B., Charles, P., & Hilberg, O. (2016). Distributed practice. The more the merrier? A randomised bronchoscopy simulation study. *Medical Education Online, 21*(1), 30517.

Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: Beliefs, techniques, and illusions. *Annual Review of Psychology, 64*(1), 417-444.

Blazek, M. C., Dantz, B., Wright, M. C., & Fiedorowicz, J. G. (2016). Spaced learning using emails to integrate psychiatry into general medical curriculum: Keep psychiatry in mind. *Medical Teacher, 38*(10), 1049-1055.

Boespflug, A., Guerra, J., Dalle, S., & Thomas, L. (2015). Enhancement of customary dermoscopy education with spaced education e-learning: a prospective controlled trial. *JAMA Dermatology, 151*(8), 847-853.

Boettcher, M., Boettcher, J., Mietzsch, S., Krebs, T., Bergholz, R., & Reinshagen, K. (2018). The spaced learning concept significantly improves training for laparoscopic suturing: a pilot randomized controlled study. *Surgical Endoscopy, 32*(1), 154-159.

Bokhorst, F. D. (1986). Confidence-weighting and the validity of achievement tests. *Psychological Reports, 59*(2), 383-386.

Bonwell, C. C., & Eison, J. A. (1991). Active Learning: Creating Excitement in the Classroom. 1991 ASHE-ERIC Higher Education Reports. ERIC Clearinghouse on Higher Education, Washington DC.

Bordage, G. (2009). Conceptual frameworks to illuminate and magnify. *Medical Education, 43*(4), 312-319.

Boron, W. F. and Boupaep, E. L. (2016) Medical Physiology. 3rd Edition, Elsevier Publisher, Philadelphia.

Botvinick, M. M. (2007). Conflict monitoring and decision making: reconciling two perspectives on anterior cingulate function. *Cognitive, Affective, & Behavioral Neuroscience, 7*(4), 356-366.

Boulet, J. R., & Durning, S. J. (2019). What we measure… and what we should measure in medical education. *Medical Education, 53*(1), 86-94.

Bowers, J. S. (2016). The practical and principled problems with educational neuroscience. *Psychological Review 123*(5), 600-612.

Braasch, J. L., Goldman, S. R., & Wiley, J. (2013). The influences of text and reader characteristics on learning from refutations in science texts. *Journal of Educational Psychology, 105*(3), 561-578.

Brandriet, A. R., & Bretz, S. L. (2014). The development of the redox concept inventory as a measure of students' symbolic and particulate redox understandings and confidence. *Journal of Chemical Education, 91*(8), 1132- 1144.

Bransen, D., Govaerts, M. J., Sluijsmans, D. M., & Driessen, E. W. (2019). Beyond the self: The role of co- regulation in medical students' self-regulated learning. *Medical Education, 54*(3), 234-241.

Bransford, J. (1979). Human cognition: Learning, understanding, and remembering. Belmont, CA: Wadsworth.

Bransford, J. D., & Johnson, M. K. (1972). Contextual prerequisites for understanding: Some investigations of comprehension and recall. *Journal of Verbal Learning and Verbal Behavior, 11*(6), 717-726.

Breckwoldt, J., Ludwig, J. R., Plener, J., Schr□der, T., Gruber, H., & Peters, H. (2016). Differences in procedural knowledge after a "spaced" and a "massed" version of an intensive course in emergency medicine, investigating a very short spacing interval. *BMC Medical Education, 16*(1), 249-257.

Brinkman, D. J., Tichelaar, J., van Agtmael, M. A., de Vries, T. P., & Richir, M. C. (2015). Self-reported confidence in prescribing skills correlates poorly with assessed competence in fourth-year medical students. J*ournal of Clinical Pharmacology, 55*(7), 825-830.

Brooks, B. J., & Koretsky, M. D. (2011). The influence of group discussion on students' responses and confidence during peer instruction. *Journal of Chemical Education, 88*(11), 1477-1484.

Brooks, J., McCluskey, S., Turley, E., & King, N. (2015). The utility of template analysis in qualitative psychology research. *Qualitative Research in Psychology, 12*(2), 202-222.

Brown, D. (2017). An evidence-based analysis of learning practices: the need for pharmacy students to employ more effective study strategies. *Currents in Pharmacy Teaching and Learning, 9*(2), 163-170.

Bruckel, J., Carballo, V., Kalibatas, O., Soule, M., Wynne, K. E., Ryan, M. P., & Shaw, T. (2016). Use of spaced education to deliver a curriculum in quality, safety and value for postgraduate medical trainees: trainee satisfaction and knowledge. *Postgraduate Medical Journal, 92*(1085), 137-144.

de Bruin, A. B. (2016). The potential of neuroscience for health sciences education: towards convergence of evidence and resisting seductive allure. *Advances in Health Sciences Education, 21*(5), 983-990.

de Bruin, A. B., Dunlosky, J., & Cavalcanti, R. B. (2017). Monitoring and regulation of learning in medical education: the need for predictive cues. *Medical Education, 51*(6), 575-584.

Brunoni, A. R., & Vanderhasselt, M.-A. (2014). Working memory improvement with non-invasive brain stimulation of the dorsolateral prefrontal cortex: a systematic review and meta-analysis. *Brain and Cognition, 86*, 1-9.

Brydges R., & Butler D. (2012). A reflective analysis of medical education research on self-regulation in learning and practice. *Medical Education, 46*(1), 71-79.

Brydges, R., Manzone, J., Shanks, D., Hatala, R., Hamstra, S. J., Zendejas, B., & Cook, D. A. (2015). Self-regulated learning in simulation-based training: a systematic review and meta-analysis. *Medical Education, 49*(4), 368- 378.

Bude, L., Imbos, T., van de Wiel, M. W., & Berger, M. P. (2011). The effect of distributed practice on students' conceptual understanding of statistics. *Higher Education, 62*(1), 69-79.

Burdo, J., & O'Dwyer, L. (2015). The effectiveness of concept mapping and retrieval practice as learning strategies in an undergraduate physiology course. *Advances in Physiology Education, 39*(4), 335-340.

Butler, A. C., Fazio, L. K., & Marsh, E. J. (2011). The hypercorrection effect persists over a week, but high- confidence errors return. *Psychonomic Bulletin & Review, 18*(6), 1238-1244.

Butterfield, B., & Metcalfe, J. (2001). Errors committed with high confidence are hypercorrected. Journal of Experimental Psychology: *Learning, Memory, and Cognition, 27*(6), 1491-1494.

## C

Cabeza, R. (2008). Role of parietal regions in episodic memory retrieval: the dual attentional processes hypothesis. *Neuropsychologia, 46*(7), 1813-1827.

Cairo, T. A., Liddle, P. F., Woodward, T. S., & Ngan, E. T. (2004). The influence of working memory load on phase specific patterns of cortical activity. *Cognitive Brain Research, 21*(3), 377-387.

Caleon, I. S., & Subramaniam, R. (2010a). Do students know what they know and what they don't know? Using a four-tier diagnostic test to assess the nature of students' alternative conceptions. *Research in Science Education, 40*(3), 313-337.

Caleon, I. S., & Subramaniam, R. (2010b). Development and application of a three-tier diagnostic test to assess secondary students' understanding of waves. *International Journal of Science Education, 32*(7), 939-961.

Canas, A., Juncadella, M., Lau, R., Gabarros, A., & Hernandez, M. (2018). Working memory deficits after lesions involving the supplementary motor area. *Frontiers in Psychology, 9*, 765.

Cannon-Bowers, J. A., Bowers, C., & Procci, K. (2010). Optimizing learning in surgical simulations: guidelines from the science of learning and human performance. *Surgical Clinics, 90*(3), 583-603.

Cao, G., & Harris, K. M. (2014). Augmenting saturated LTP by broadly spaced episodes of theta-burst stimulation in hippocampal area CA1 of adult rats and mice. J*ournal of Neurophysiology, 112*(8), 1916-1924.

Carew, T. J. & Magsamen, S. H. (2010). Neuroscience and education: An ideal partnership for producing evidence-based solutions to guide 21st century learning. *Neuron 67*(5), 685-688.

Carpenter, S. K., Cepeda, N. J., Rohrer, D., Kang, S. H., & Pashler, H. (2012). Using spacing to enhance diverse forms of learning: review of recent research and implications for instruction. *Educational Psychology Review, 24*(3), 369–378.

Carpenter, S. K., Haynes, C. L., Corral, D., & Yeung, K. L. (2018). Hypercorrection of high-confidence errors in the classroom. *Memory, 26*(10), 1379-1384.

Cavanna, A. E., & Trimble, M. R. (2006). The precuneus: a review of its functional anatomy and behavioural correlates. *Brain, 129*(3), 564-583.

Cavers M., & Ling J. (2016). Confidence Weighting Procedures for MultipleChoice Tests. In: Chen D. G., Chen J., Lu X., Yi G., Yu H., editors. Advanced Statistical Methods in Data Science. ICSA Book Series in Statistics. Singapore: Springer.

Cecilio-Fernandes, D., Cnossen, F., Jaarsma, D. A., & Tio, R. A. (2018). Avoiding surgical skill decay: a systematic review on the spacing of training sessions. *Journal of Surgical Education, 75*(2), 471-480.

Cepeda, N. J., Coburn, N., Rohrer, D., Wixted, J. T., Mozer, M. C., & Pashler, H. (2009). Optimizing distributed practice: theoretical analysis and practical implications. *Experimental Psychology, 56*(4), 236-246.

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*(11), 1095-1102.

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin, 132*(3), 354.

Chamberland, M., & Mamede, S. (2015). Self-explanation, an instructional strategy to foster clinical reasoning in medical students. *Health Professions Education, 1*(1), 24-33.

Chang, C., Crottaz-Herbette, S., & Menon, V. (2007). Temporal dynamics of basal ganglia response and connectivity during verbal working memory. *Neuroimage, 34*(3), 1253-1269.

Chen, O., Manalo, E., & She, Y. (2019). Examining the influence of expertise on the effectiveness of diagramming and summarising when studying scientific materials. *Educational Studies, 45*(1), 57-71.

Chi, M. T. (2005). Commonsense conceptions of emergent processes: Why some misconceptions are robust. *The Journal of the Learning Sciences, 14*(2), 161-199.

Chi, M. T. (2009a). Active-constructive-interactive: A conceptual framework for differentiating learning activities. *Topics in Cognitive Science, 1*(1), 73-105.

Chi, M. T. (2009b). Three types of conceptual change: Belief revision, mental model transformation, and categorical shift. In International handbook of research on conceptual change (pp. 89-110). NY: Routledge.

Chi, M. T., & Bassok, M. (1989). Learning from examples via self-explanations. In Knowing, learning, and instruction: Essays in honor of Robert Glaser (pp. 251-282). NJ: Lawrence Erlbaum Associates.

Chi, M. T., & Roscoe, R. D. (2002). The processes and challenges of conceptual change. In Reconsidering conceptual change: Issues in theory and practice (pp. 3-27). Dordrecht: Springer.

Chi, M. T., Slotta, J. D., & De Leeuw, N. (1994a). From things to processes: A theory of conceptual change for learning science concepts. *Learning and Instruction, 4*(1), 27-43.

Chi, M. T., Leeuw, N., Chiu, M. H., & LaVancher, C. (1994b). Eliciting self-explanations improves understanding. *Cognitive Science, 18*(3), 439-477.

Chi, M. T., Roscoe, R. D., Slotta, J. D., Roy, M., & Chase, C. C. (2012). Misconceived causal explanations for emergent processes. *Cognitive Science, 36*(1), 1-61.

Chi, M. T., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist, 49*(4), 219-243.

Chua, E. F., Pergolizzi, D., & Weintraub, R. R. (2014). The cognitive neuroscience of metamemory monitoring: understanding metamemory processes, subjective levels expressed, and metacognitive accuracy. In: The cognitive neuroscience of metacognition (pp. 267-291). Dordrecht: Springer.

Cilliers, F. J. (2015). Is assessment good for learning or learning good for assessment? A. Both? B. Neither? C. It depends?. *Perspectives on Medical Education, 4*(6), 280-281.

Cleary, T. J., Durning, S. J., & Artino, A. R. (2016). Microanalytic assessment of self-regulated learning during clinical reasoning tasks: recent developments and next steps. *Academic Medicine, 91*(11), 1516-1521.

Collier, A. (1994). Critical realism: an introduction to Roy Bhaskar's philosophy. London: Verso.

Colman, A.M. (2001). The Dictionary of Psychology. Oxford, UK: Oxford University Press.

Connor, E. V., Raker, C., & Wohlrab, K. J. (2016). Effects of repetition and inactivity on laparoscopic skills training. *Journal of Minimally Invasive Gynecology, 23*(2), 194-197.

Cordova, J. R., Sinatra, G. M., Jones, S. H., Taasoobshirazi, G., & Lombardi, D. (2014). Confidence in prior knowledge, self-efficacy, interest and prior knowledge: Influences on conceptual change. *Contemporary Educational Psychology, 39*(2), 164-174.

Cortright, R. N., Collins, H. L., & DiCarlo, S. E. (2005). Peer instruction enhanced meaningful learning: Ability to solve novel problems. *Advances in Physiology Education, 29*(2), 107-111.

Croskerry, P., & Norman, G. (2008). Overconfidence in clinical decision making. *The American Journal of Medicine, 121*(5), S24-S29.

Crouch, C. H., & Mazur, E. (2001). Peer Instruction: Ten years of experience and results. *American Journal of Physics, 69*(9), 970-977.

Curtis, D. A., Lind, S. L., Boscardin, C. K., & Dellinges, M. (2013). Does student confidence on multiple-choice question assessments provide useful information? *Medical Education 47*(6), 578-584.

Curtis, D. A., Lind, S. L., Dellinges, M., & Schroeder, K. (2012). Identifying student misconceptions in biomedical course assessments in dental education. *Journal of Dental Education, 76*(9), 1183-1194.

Custers, E. J., & ten Cate, O. T. (2011). Very long-term retention of basic science knowledge in doctors after graduation. *Medical Education, 45*(4), 422-430.

## D

Daniel, T. A., Katz, J. S., & Robinson, J. L. (2016). Delayed match-to-sample in working memory: A BrainMap meta-analysis. *Biological Psychology, 120*(1), 10-20.

Dawson-Saunders B., Feltovich, P. J., Coulson, R. L., Steward, D. E. (1990). A survey of medical school teachers to identify basic biomedical concepts medical students should understand. *Academic Medicine. 65*(7), 448-454.

Deng, F., Gluckstein, J. A., & Larsen, D. P. (2015). Student-directed retrieval practice is a predictor of medical licensing examination performance. *Perspectives on Medical Education, 4*(6), 308-313.

Diakidoy, I. A. N., Mouskounti, T., Fella, A., & Ioannides, C. (2016). Comprehension processes and outcomes with refutation and expository texts and their contribution to learning. *Learning and Instruction, 41*, 60-69.

Dienes Z., & Scott R. (2005). Measuring unconscious knowledge: Distinguishing structural knowledge and judgment knowledge. *Psychological Research, 69*(5-6), 338-351.

Dienes Z., & Seth A. (2010). Gambling on the unconscious: A comparison of wagering and confidence ratings as measures of awareness in an artificial grammar task. *Consciousness and Cognition, 19*(2), 674-681.

Dobson, J. L. (2011). Effect of selected "desirable difficulty" learning strategies on the retention of physiology information. *Advances in Physiology Education, 35*(4), 378-383.

Dobson, J. L. (2012). Effect of uniform versus expanding retrieval practice on the recall of physiology information. *Advances in Physiology Education, 36*(1), 6-12.

Dobson, J. L. (2013). Retrieval practice is an efficient method of enhancing the retention of anatomy and physiology information. *Advances in Physiology Education, 37*(2), 184-191.

Dobson, J. L., Perez, J., & Linderholm, T. (2017). Distributed retrieval practice promotes superior recall of anatomy information. *Anatomical Sciences Education, 10*(4), 339-347.

Dougherty, M. R. & Robey, A. (2018). Neuroscience and Education: A Bridge Astray? *Current Directions in Psychological Science 27*(6), 401-406.

Duit, R., & Treagust, D. F. (1995). Students' conceptions and constructivist teaching approaches. In: Improving Science Education, edited by Fraser BJ, Walberg HJ. (pp. 46-49). Chicago, IL: University of Chicago Press.

Duit, R., & Treagust, D. F. (2012). How can conceptual change contribute to theory and practice in science education? In: Second International Handbook of Science Education, edited by Fraser B, Tobin K, McRobbie C. (pp. 107-118). Dordrecht: Springer.

Dunbar, K. N., Fugelsang, J. A., & Stein, C. (2007). Do Naive Theories Ever Go Away? Using Brain and Behavior to Understand Changes in Concepts. In: Thinking with data (pp. 205-217). London: Psychology Press.

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques. *Psychological Science in the Public Interest, 14*(1), 4-58.

Durning, S. J., Dolmans, D. H., Cleland, J., Mennin, S., Amin, Z., & Gibbs, T. J. (2012). The AMEE Research Committee: Initiatives to stimulate research and practice. *Medical Teacher, 34*(6), 458-461.

# E

Ebbinghaus, H. (1885/1913). Memory: A Contribution to Experimental Psychology. NY: Teachers College, Columbia University.

Ecker, U. K., Lewandowsky, S., & Tang, D. T. (2010). Explicit warnings reduce but do not eliminate the continued influence of misinformation. *Memory and Cognition, 38*(8), 1087-1100.

Ecker, U. K., Lewandowsky, S., Swire, B., & Chang, D. (2011). Correcting false information in memory: manipulating the strength of misinformation encoding and its retraction. *Psychonomic Bulletin and Review, 18*(3), 570-578.

Ell, S. W., Helie, S., Hutchinson, S., Costa, A., & Villalba, E. (2011). Contributions of the putamen to cognitive function. (pp. 29-52). NY: Nova Science Publishers Hauppauge.

Ernst, K. D., Cline, W. L., Dannaway, D. C., Davis, E. M., Anderson, M. P., Atchley, C. B., & Thompson, B. M. (2014). Weekly and consecutive day neonatal intubation training: comparable on a pediatrics clerkship. *Academic Medicine, 89*(3), 505-510.

Evans, D. J., Zeun, P., & Stanier, R. A. (2014). Motivating student learning using a formative assessment journey. *Journal of Anatomy, 224*(3), 296-303.

# F

Fann, J. I., Caffarelli, A. D., Georgette, G., Howard, S. K., Gaba, D. M., Youngblood, P., ... & Burdon, T. A. (2008). Improvement in coronary anastomosis with cardiac surgery simulation. *The Journal of Thoracic and Cardiovascular Surgery, 136*(6), 1486-1491.

Farah, C. A., Weatherill, D., Dunn, T. W., & Sossin, W. S. (2009). PKC differentially translocates during spaced and massed training in Aplysia. *Journal of Neuroscience, 29*(33), 10281-10286.

Farley, J., Risko, E., & Kingstone, A. (2013). Everyday attention and lecture retention: the effects of time, fidgeting, and mind wandering. *Frontiers in Psychology, 4*(4), 619.

Favazzo, L., Willford, J. D., & Watson, R. M. (2014). Correlating student knowledge and confidence using a graded knowledge survey to assess student learning in a general microbiology classroom. *Journal of Microbiology and Biology Education, 15*(2), 251.

Feltovich, P. J., Spiro, P. J., & Coulson, R. L. (1988). The nature of conceptual understanding in biomedicine: The deep structure of complex ideas and the development of misconceptions. *Center for the Study of Reading Technical Report*; no. 440.

Feltovich, P. J., Spiro, R. J., & Coulson, R. L. (1993). Learning, teaching, and testing for complex conceptual understanding. In: Test theory for a new generation of tests (pp. 181-217). NJ: Lawrence Erlbaum Associates.

Fields, R. D. (2005). Making memories stick. *Scientific American, 292*(2), 74-81.

Finnerty, E. P., Chauvin, S., Bonaminio, G., Andrews, M., Carroll, R. G., & Pangaro, L. N. (2010). Flexner revisited: The role and value of the basic sciences in medical education. *Academic Medicine, 85*(2), 349-355.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive– developmental inquiry. *American Psychologist, 34*(10), 906-911.

Fleming, S. M., Dolan, R. J. (2010). Effects of loss aversion on post-decision wagering: implications for measures of awareness. *Consciousness and Cognition, 19*(1), 352-363.

Fleming, S. M., Ryu, J., Golfinos, J. G., & Blackmon, K. E. (2014). Domain-specific impairment in metacognitive accuracy following anterior prefrontal lesions. *Brain, 137*(10), 2811-2822.

Flett, K. B., Bousvaros, A., Carpenter, J., Millrinen, C. E., Martin, P., & Sandora, T. J. (2018). Reducing redundant anaerobic therapy through spaced education and antimicrobial stewardship interventions. *Journal of the Pediatric Infectious Diseases Society, 7*(4), 317-322.

Flobakk, F. R. (2015). The development and impact of educational neuroscience. A critical discourse analysis. Trondheim: Norwegian University of Science and Technology.

Foisy, L.-M. B., Potvin, P., Riopel, M., & Masson, S. (2015). Is inhibition involved in overcoming a common physics misconception in mechanics? *Trends in Neuroscience and Education, 4*(1-2), 26-36.

Freda, N. M., & Lipp, M. J. (2016). Test-enhanced learning in competence-based predoctoral orthodontics: a four- year study. *Journal of Dental Education, 80*(3), 348-354.

Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., & Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences of the USA, 111*(23), 8410-8415.

Friedlander, M. J., Andrews, L., Armstrong, E. G., Aschenbrenner, C., Kass, J. S., Ogden, P., Schwartzstein, R., & Viggiano, T. R. (2011). What can medical education learn from the neurobiology of learning? *Academic Medicine, 86*(4), 415-420.

Friedman, C. P., Gatti, G. G., Franz, T. M., Murphy, G. C., Wolf, F. M., Heckerling, P. S., Fine, P. L., Miller, T. M., & Elstein, A. S. (2005). Do physicians know when their diagnoses are correct? Implications for decision support and error reduction. *Journal of General Internal Medicine, 20*(4), 334-339.

Fugelsang, J. A., & Dunbar, K. N. (2005). Brain-based mechanisms underlying complex causal thinking.

*Neuropsychologia, 43*(8), 1204-1213.

Fyrenius, A., Silen, C., & Wirell, S. (2007). Students' conceptions of underlying principles in medical physiology: an interview study of medical students' understanding in a PBL curriculum. *Advances in Physiology Education, 31*(4), 364-369.

## G

Gabrieli, J. D. (2016). The promise of educational neuroscience: Comment on Bowers (2016). *Psychological Review, 123*(5), 613-619.

Gall, M. D. (1990). Tools for Learning: A Guide to Teaching Study Skills. Association for Supervision and Curriculum Development. VA: Alexandria.

Galvagno Jr, S. M., & Segal, B. S. (2009). Critical action procedures testing: a novel method for test-enhanced learning. *Medical Education, 43*(12), 1182-1187.

Gandhi, M., Beasley, A., Vinas, E., Sangi-Haghpeykar, H., Ramin, S. M., & Kilpatrick, C. C. (2016). Electronic Learning–Spaced Education to Facilitate Resident Knowledge and Guide Program Didactics. *Obstetrics & Gynecology, 128*, 23S-26S.

Gandomkar, R., Mirzazadeh, A., Jalili, M., Yazdani, K., Fata, L., & Sandars, J. (2016). Self-regulated learning processes of medical students during an academic learning task. *Medical Education, 50*(10), 1065-1074.

Garzia, M., Mangione, G. R., Longo, L., & Pettenati, M. C. (2016). Spaced learning and innovative teaching: school time, pedagogy of attention and learning awareness. *Research on Education and Media, 8*(1), 22-37.

Genoux, D., Haditsch, U., Knobloch, M., Michalon, A., Storm, D., & Mansuy, I. M. (2002). Protein phosphatase 1 is a molecular constraint on learning and memory. *Nature, 418*(6901), 970-975.

Giles, R. M., Johnson, M. R., Knight, K. E., Zammett, S., & Weinman, J. (1982). Recall of lecture information: a question of what, when and where. *Medical Education, 16*(5), 264-268.

Gill, D. L. (2011). Beyond the qualitative–quantitative dichotomy: notes from a non-qualitative researcher. *Qualitative Research in Sport, Exercise and Health, 3*(3), 305-312.

Giuliodori, M. J., Lujan, H. L., & DiCarlo, S. E. (2006). Peer instruction enhanced student performance on qualitative problem-solving questions. *Advances in Physiology Education, 30*(4), 168-173.

Gluckman, M., Vlach, H. A., & Sandhofer, C. M. (2014). Spacing simultaneously promotes multiple forms of learning in children's science curriculum. *Applied Cognitive Psychology, 28*(2), 266-273.

Goldszmidt, M., Minda, J. P., Devantier, S. L., Skye, A. L., & Woods, N. N. (2012). Expanding the basic science debate: The role of physics knowledge in interpreting clinical findings. *Advances in Health Sciences Education, 17*(4), 547-555.

Gooding, H., Mann, K., & Armstrong, E. (2017). Twelve tips for applying the science of learning to health professions education. *Medical Teacher, 39*(1), 26-31.

Goswami, U. (2006). Neuroscience and education: from research to practice? *Nature Reviews*

*Neuroscience, 7*(5), 406- 411.

Graesser, A. C., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review, 101*(3), 371-395.

Grande, J. P. (2009). Training of physicians for the twenty-first century: role of the basic sciences. *Medical Teacher, 31*(9), 802-806.

Grazziotin-Soares, R., Lind, S. L., Ardenghi, D. M., & Curtis, D. A. (2018). Misconceptions amongst dental students: How can they be identified? *European Journal of Dental Education, 22*(1), e101-e106.

Group, S. N. W. (1996). Skills for the new millennium: CanMEDS 2000 project. *Annals Royal College of Physicians and Surgeons of Canada, 29*, 206-216.

Gunstone R.F., & Mitchell, I.J. (2005). Metacognition and conceptual change. In: Mintzes JJ, Wandersee JH, Novak JD, editors. Teaching science for understanding. A human constructivist view. (pp. 133-163). San Diego, CA: Academic Press.

Guzzetti, B. J., Snyder, T. E., Glass, G. V., & Gamas, W. S. (1993). Promoting conceptual change in science: A comparative meta-analysis of instructional interventions from reading education and science education. *Reading Research Quarterly, 28*, 117-159.

Gyorki, D. E., Shaw, T., Nicholson, J., Baker, C., Pitcher, M., Skandarajah, A., ... & Mann, G. B. (2013). Improving the impact of didactic resident training with online spaced education. *ANZ Journal Of Surgery, 83*(6), 477-480.

## H

Hailikari, T., Katajavuori, N., & Lindblom-Ylanne, S. (2008). The relevance of prior knowledge in learning and instructional design. *American Journal of Pharmaceutical Education, 72*(5), 113.

Halliday, N., O'Donoghue, D., Klump, K. E., & Thompson, B. (2015). Human structure in six and one-half weeks: One approach to providing foundational anatomical competency in an era of compressed medical school anatomy curricula. *Anatomical Sciences Education, 8*(2), 149-157.

Hardiman, M., Rinne, L., Gregory, E. & Yarmolinskaya, J. (2012). Neuroethics, neuroeducation, and classroom teaching: Where the brain sciences meet pedagogy. *Neuroethics 5*(2), 135-143.

Hartman, H. J. (2001). Developing students' metacognitive knowledge and skills. In: Metacognition in learning and instruction (pp. 33-68). Dordrecht: Springer.

Hasan, S., Bagayoko, D., & Kelley, E. L. (1999). Misconceptions and the certainty of response index (CRI). *Physics Education, 34*(5), 294-299.

Hernick, M. (2015). Test-enhanced learning in an immunology and infectious disease medicinal chemistry/pharmacology course. *American Journal of Pharmaceutical Education, 79*(7), 97.

Herwaarden, C. L. A., Laan, R. F. J. M., & Leunissen, R. R. M. (Eds.). (2009). Raamplan artsopleiding 2009. Nederlandse Federatie van Universitair Medische Centra (NFU).

Hewson, M. G., & Hewson, P. W. (1983). Effect of instruction using students' prior knowledge and conceptual change strategies on science learning. *Journal of Research in Science Teaching, 20*(8), 731-743.

van Hoof, T. J, & Doyle, T. J. (2018). Learning science as a potential new source of understanding and improvement for continuing education and continuing professional development. *Medical Teacher, 40*(9), 880-885.

Hopkins, K.D., Hakstian, A.R., Hopkins, B.R. (1973). Validity and reliability consequences of confidence weighting. *Educational and Psychological Measurement. 33*(1), 135-41.

Horn, S., & Hernick, M. (2015). Improving student understanding of lipids concepts in a biochemistry course using test-enhanced learning. *Chemistry Education Research and Practice, 16*(4), 918-928.

van Houten-Schat, M. A., Berkhout, J. J., van Dijk, N., Endedijk, M. D., Jaarsma, A. D. C., & Diemers, A. D. (2018). Self-regulated learning in the clinical context: a systematic review. *Medical Education, 52*(10), 1008-1015.

Howard-Jones, P. (2009). Introducing neuroeducational research: Neuroscience, education and the brain from contexts to practice. NY: Routledge.

Howard-Jones, P. A., Varma, S., Ansari, D., Butterworth, B., De Smedt, B., Goswami, U., Laurillard, D., & Thomas, M. S. (2016). The principles and practices of educational neuroscience: Comment on Bowers (2016). *Psychological Review, 123*(5), 620-627.

Hsieh, H. F., & Shannon, S. E. (2005). Three approaches to qualitative content analysis. *Qualitative Health Research, 15*(9), 1277-1288.

Hung, Y., Gaillard, S. L., Yarmak, P., & Arsalidou, M. (2018). Dissociations of cognitive inhibition, response inhibition, and emotional interference: Voxelwise ALE meta-analyses of fMRI studies. *Human Brain Mapping, 39*(10), 4065-4082.

Hunt, D. P. (2003). The concept of knowledge and how to measure it. *Journal of Intellectual Capital, 4*(1), 100-113.

# I

-

# J

Jackson, T. H., Hannum, W. H., Koroluk, L., & Proffit, W. R. (2011). Effectiveness of web-based teaching modules: test-enhanced learning in dental education. *Journal of Dental Education, 75*(6), 775-781.

James, M. C., & Willoughby, S. (2011). Listening to student conversations during clicker questions: What you have not heard might surprise you! *American Journal of Physics, 79*(1), 123-132.

de Jong, T., Van Gog, T., Jenks, K., Manlove, S., Van Hell, J., Jolles, J., & Boschloo, A. (2009). Explorations in learning and the brain: On the potential of cognitive neuroscience for educational science. NY: Springer Science & Business Media.

# K

Kampmeyer, D., Matthes, J., & Herzig, S. (2015). Lucky guess or knowledge: a cross-sectional study using the Bland and Altman analysis to compare confidence-based testing of pharmacological knowledge in 3rd and 5th year medical students. *Advances in Health Sciences Education, 20*(2), 431-440.

Kandel, E. R., Dudai, Y., & Mayford, M. R. (2014). The molecular and systems biology of memory. *Cell, 157*(1), 163- 186.

Karpicke, J. D., & Roediger, H. L. (2008). The critical importance of retrieval for learning. *Science, 319*(5865), 966- 968.

Katz, L. M., McKinnish, T., Gilliland, K., Tolleson-Rinehart, S., & Marks, B. L. (2017). Teaching procedural skills to medical students: a pilot procedural skills lab. *Education for Health, 30*(1), 79-83.

Keene, A. B., Shiloh, A. L., Dudaie, R., Eisen, L. A., & Savel, R. H. (2012). Online testing from Google Docs™ to enhance teaching of core topics in critical care: a pilot study. *Medical Teacher, 34*(12), 1075-1077.

Kelemen, D., & Rosset, E. (2009). The human function compunction: Teleological explanation in adults. *Cognition, 111*(1), 138-143.

Kelemen, D., Rottman, J., & Seston, R. (2013). Professional physical scientists display tenacious teleological tendencies: Purpose-based reasoning as a cognitive default. *Journal of Experimental Psychology: General, 142*(4), 1074.

Kelley, P., & Whatson, T. (2013). Making long-term memories in minutes: a spaced learning pattern from memory research in education. *Frontiers in Human Neuroscience, 7*(589), 1-9.

Kendeou, P., & O'Brien, E. J. (2014). The Knowledge Revision Components (KReC) Framework: Processes and Mechanisms. Processing inaccurate information: *T*heoretical and Applied Perspectives From Cognitive Science and the Educational Sciences (pp. 353-377). Cambridge, MA: MIT Press.

Kendeou, P., & Van Den Broek, P. (2007). The effects of prior knowledge and text structure on comprehension processes during reading of scientific texts. *Memory & Cognition, 35*(7), 1567-1577.

Kerdijk, W., Cohen-Schotanus, J., Mulder, B. F., Muntinghe, F. L., & Tio, R. A. (2015). Cumulative versus end-of- course assessment: effects on self-study time and test performance. *Medical Education, 49*(7), 709-716.

Kerdijk, W., Tio, R. A., Mulder, B. F., & Cohen-Schotanus, J. (2013). Cumulative assessment: strategic choices to influence students' study effort. *BMC Medical Education, 13*(1), 172.

Kerfoot, B. P. (2010). Adaptive spaced education improves learning efficiency: a randomized controlled trial. *The Journal of Urology, 183*(2), 678-681.

Kerfoot, B. P. (2009). Learning benefits of on-line spaced education persist for 2 years. *The Journal of*

*Urology, 181*(6), 2671-2673.

Kerfoot, B. P. (2008). Interactive spaced education versus web based modules for teaching urology to medical students: a randomized controlled trial. *The Journal of Urology, 179*(6), 2351-2357.

Kerfoot, B. P., Armstrong, E. G., & O'Sullivan, P. N. (2008). Interactive spaced-education to teach the physical examination: a randomized controlled trial. *Journal of General Internal Medicine, 23*(7), 973-978.

Kerfoot, B. P., & Baker, H. (2012a). An online spaced-education game for global continuing medical education: a randomized trial. *Annals of Surgery, 256*(1), 33-38.

Kerfoot, B. P., & Baker, H. (2012b). An online spaced-education game to teach and assess residents: a multi- institutional prospective trial. *Journal of the American College of Surgeons, 214*(3), 367-373.

Kerfoot, B. P., Baker, H., Pangaro, L., Agarwal, K., Taffet, G., Mechaber, A. J., & Armstrong, E. G. (2012). An online spaced-education game to teach and assess medical students: a multi-institutional prospective trial. *Academic Medicine, 87*(10), 1443-1449.

Kerfoot, B. P., Baker, H. E., Koch, M. O., Connelly, D., Joseph, D. B., & Ritchey, M. L. (2007a). Randomized, controlled trial of spaced education to urology residents in the United States and Canada. *The Journal of Urology, 177*(4), 1481-1487.

Kerfoot, B. P., & Brotschi, E. (2009). Online spaced education to teach urology to medical students: a multi- institutional randomized trial. *The American Journal of Surgery, 197*(1), 89-95.

Kerfoot, B. P., DeWolf, W. C., Masser, B. A., Church, P. A., & Federman, D. D. (2007b). Spaced education improves the retention of clinical knowledge by medical students: a randomised controlled trial. *Medical Education, 41*(1), 23-31.

Kerfoot, B. P., Fu, Y., Baker, H., Connelly, D., Ritchey, M. L., & Genega, E. M. (2010a). Online spaced education generates transfer and improves long-term retention of diagnostic skills: a randomized controlled trial. *Journal of the American College of Surgeons, 211*(3), 331-337.

Kerfoot, B. P., Kearney, M. C., Connelly, D., & Ritchey, M. L. (2009). Interactive spaced education to assess and improve knowledge of clinical practice guidelines: a randomized controlled trial. *Annals of Surgery, 249*(5), 744-749.

Kerfoot, B. P., Lawler, E. V., Sokolovskaya, G., Gagnon, D., & Conlin, P. R. (2010b). Durable improvements in prostate cancer screening from online spaced education: a randomized controlled trial. *American Journal of Preventive Medicine, 39*(5), 472-478.

Kerfoot, B. P., Shaffer, K., McMahon, G. T., Baker, H., Kirdar, J., Kanter, S., ... & Armstrong, E. G. (2011). Online "spaced education progress-testing" of students to confront two upcoming challenges to medical schools. *Academic Medicine, 86*(3), 300-306.

Kerfoot, B. P., Turchin, A., Breydo, E., Gagnon, D., & Conlin, P. R. (2014). An online spaced-education game among clinicians improves their patients' time to blood pressure control: a randomized controlled trial. *Circulation: Cardiovascular Quality and Outcomes, 7*(3), 468-474.

Kerfoot, B. P., Armstrong, E. G., & O'Sullivan, P. N. (2008). Impact of item clustering on interactive spaced education. *Medical Education, 42*(11), 1115-1116.

Kesser, B. W., Hallman, M., Murphy, L., Tillar, M., Keeley, M., & Peirce, S. (2014). Interval vs massed training: how best do we teach surgery?. *Otolaryngology-Head and Neck Surgery, 150*(1), 61-67.

van Kesteren, M.T., Ruiter, D.J., Fernandez, G., & Henson, R.N. (2012). How schema and novelty augment memory formation. *Trends in Neurosciences. 35*(2), 211-219.

Kim, H. (2010). Dissociating the roles of the default-mode, dorsal, and ventral networks in episodic memory retrieval. *Neuroimage, 50*(4), 1648-1657.

King, A. (1992). Comparison of self-questioning, summarizing, and notetaking-review as strategies for learning from lectures. *American Educational Research Journal, 29*(2), 303-323.

Kintsch, W. (1998). Comprehension: A paradigm for cognition. NY: Cambridge University Press.

Kleiman, A. M., Forkin, K. T., Bechtel, A. J., Collins, S. R., Ma, J. Z., Nemergut, E. C., & Huffmyer, J. L. (2017). Generative retrieval improves learning and retention of cardiac anatomy using transesophageal echocardiography. *Anesthesia & Analgesia, 124*(5), 1440-1444.

Klymkowsky, M. W., Taylor, L. B., Spindler, S. R., & Garvin-Doxas, R. K. (2006). Twodimensional, implicit confidence tests as a tool for recognizing student misconceptions. *Journal of College Science Teaching, 36*(3), 44-48.

Koch, C., & Preuschoff, K. (2017). Betting the house on consciousness. *Nature Neuroscience 10*(2), 140-141.

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

Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. *Theory into Practice, 41*(4), 212-218.

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology, 77*(6), 1121-1134.

Kulasegaram, K. M., Chaudhary, Z., Woods, N., Dore, K., Neville, A., & Norman, G. (2017). Contexts, concepts and cognition: Principles for the transfer of basic science knowledge. *Medical Education, 51*(2), 184-195.

Kulasegaram, K. M., Martimianakis, M. A., Mylopoulos, M., Whitehead, C. R., & Woods, N. N. (2013). Cognition before curriculum: Rethinking the integration of basic science and clinical learning. *Academic Medicine, 88*(10), 1578-1585.

Kulasegaram, K., M., Ames, K., Howey, E., Neville, A., & Norman, G. (2012). The effect of conceptual and contextual familiarity on transfer performance. *Advances in Health Sciences Education, 17*(4), 489-499.

Kulasegaram, K., M., Howey, E., Neville, A., Woods, N., Dore, K., & Norman, G. (2015). The mediating effect of context variation in mixed practice for transfer of basic science. *Advances in Health Sciences Education, 20*(4), 953-968.

Kurosawa, H., Ikeyama, T., Achuff, P., Perkel, M., Watson, C., Monachino, A., ... & Berg, R. A. (2014). A randomized, controlled trial of in situ pediatric advanced life support recertification ("pediatric advanced life support reconstructed") compared with standard pediatric advanced life support recertification for ICU frontline providers. *Critical Care Medicine, 42*(3), 610-618.

# L

LaBossiere, M. D., Dell, K. A., Sunjic, K., & Wantuch, G. A. (2016). Student perceptions of group examinations as a method of exam review in pharmacotherapeutics. *Currents in Pharmacy Teaching and Learning, 8*(3), 375-379.

LaDisa, A. G., & Biesboer, A. (2017). Incorporation of practice testing to improve knowledge acquisition in a pharmacotherapy course. *Currents in Pharmacy Teaching and Learning, 9*(4), 660-665.

Lai, N. M., & Teng, C. L. (2011). Self-perceived competence correlates poorly with objectively measured competence in evidence based medicine among medical students. *BMC Medical Education, 11*(1), 25.

Laksov, K. B., Dornan, T., & Teunissen, P. W. (2017). Making theory explicit-An analysis of how medical education research(ers) describe how they connect to theory. BMC *Medical Education, 17*(1), 18.

Langendyk, V. (2006). Not knowing that they do not know: self-assessment accuracy of third-year medical students. *Medical Education, 40*(2), 173-179.

Larsen, D. P., Butler, A. C., Lawson, A. L., & Roediger, H. L. (2013a). The importance of seeing the patient: test- enhanced learning with standardized patients and written tests improves clinical application of knowledge. *Advances in Health Sciences Education, 18*(3), 409-425.

Larsen, D. P., Butler, A. C., & Roediger III, H. L. (2013b). Comparative effects of test-enhanced learning and self- explanation on long-term retention. *Medical Education, 47*(7), 674-682.

Larsen, D. P., Butler, A. C., Aung, W. Y., Corboy, J. R., Friedman, D. I., & Sperling, M. R. (2015). The effects of test-enhanced learning on long-term retention in AAN annual meeting courses. *Neurology, 84*(7), 748-754.

Lasry, N., Charles, E., & Whittaker, C. (2016). Effective variations of peer instruction: The effects of peer discussions, committing to an answer, and reaching a consensus. *American Journal of Physics, 84*(8), 639-645.

Legge, J. (1971). Confucian Analects, The Great Learning and The Doctrine of the Mean. NY: Dover.

de Leon, J. (2018). Teaching medical students how to think: narrative, mechanistic and mathematical thinking. *Actas Espanolas de Psiquiatria, 46*(4), 133-145.

Levac, D., Colquhoun, H., & O'Brien, K. K. (2010). Scoping studies: advancing the methodology. *Implementation Science, 5*(1), 69.

Lin, L. M., & Zabrucky, K. M. (1998). Calibration of comprehension: Research and implications for education and instruction. *Contemporary Educational Psychology, 23*(4), 345-391.

Lin, Y., & Cheng, A. (2015). The role of simulation in teaching pediatric resuscitation: current perspectives. *Advances in Medical Education and Practice, 6*, 239-248.

Lindquist, S. I., & McLean, J. P. (2011). Daydreaming and its correlates in an educational environment. *Learning and Individual Differences, 21*(2), 158-167.

Lindsey, B. A., & Nagel, M. L. (2015). Do students know what they know? Exploring the accuracy of students' self- assessments. *Physical Review Special Topics-Physics Education Research, 11*(2), 020103.

Linton, D. L., Farmer, J. K., & Peterson, E. (2014). Is peer interaction necessary for optimal active learning? *CBE Life Sciences Education, 13*(2), 243-252.

Lisk, K., Agur, A. M., & Woods, N. N. (2016). Exploring cognitive integration of basic science and its effect on diagnostic reasoning in novices. P*erspectives on Medical Education, 5*(3), 147-153.

Lisk, K., Agur, A. M. R., & Woods, N. N. (2017). Examining the effect of self-explanation on cognitive integration of basic and clinical sciences in novices. *Advances in Health Sciences Education, 22*(5), 1071-1083.

van Loon, M. H., de Bruin, A. B., van Gog, T., van Merrienboer, J. J., & Dunlosky, J. (2014). Can students evaluate their understanding of cause-and-effect relations? The effects of diagram completion on monitoring accuracy. *Acta Psychologica, 151*, 143-154.

van Loon, M. H., Dunlosky, J., Van Gog, T., Van Merrienboer, J. J., & De Bruin, A. B. (2015). Refutations in science texts lead to hypercorrection of misconceptions held with high confidence. *Contemporary Educational Psychology, 42*, 39-48.

Loyens, S. M., Jones, S. H., Mikkers, J., & van Gog, T. (2015). Problem-based learning as a facilitator of conceptual change. *Learning and Instruction, 38*, 34-42.

Lucieer, S. M., Jonker, L., Visscher, C., Rikers, R. M., & Themmen, A. P. (2016). Self-regulated learning and academic performance in medical education. *Medical Teacher, 38*(6), 585-593.

## M

Mackay, S., Morgan, P., Datta, V., Chang, A., & Darzi, A. (2002). Practice distribution in procedural skills training. *Surgical Endoscopy and Other Interventional Techniques, 16*(6), 957-961.

Macleod, C. M. (2007). The concept of inhibition in cognition. In: D. S. Gorfein & C. M. MacLeod (Eds.), Inhibition in cognition (pp. 3–23). Washington, DC: American Psychological Association.

Mahler, S. A., Wolcott, C. J., Swoboda, T. K., Wang, H., & Arnold, T. C. (2011). Techniques for teaching electrocardiogram interpretation: self-directed learning is less effective than a workshop or lecture. *Medical Education, 45*(4), 347-353.

Marcel, F. D. (2006). Knowledge loss of medical students on first year basic science courses at the University of Saskatchewan. *BMC Medical Education, 6*(1), 5.

Marcotte, D. B., Geyer, P. R., Kilpatrick, D. G., & Smith, A. D. (1976). The effect of a spaced sex education course on medical students' sexual knowledge and attitudes. *Medical Education, 10*(2), 117-121.

Mareschal, D., Butterworth, B., & Tolmie, A. (2013). Educational neuroscience: John Wiley & Sons.

Marton, F., & Saljo, R. (1976). On qualitative differences in learning: I. Outcome and process. *British Journal of Educational Psychology, 46*(1), 4-11.

Mason, L., Ariasi, N., & Boldrin, A. (2011). Epistemic beliefs in action: Spontaneous reflections about knowledge and knowing during online information searching and their influence on learning. *Learning and Instruction, 21*(1), 137-151.

Mason, L., Baldi, R., Di Ronco, S., Scrimin, S., Danielson, R. W., & Sinatra, G. M. (2017). Textual and graphical refutations: Effects on conceptual change learning. *Contemporary Educational Psychology, 49*, 275-288.

Mason, L., Gava, M., & Boldrin, A. (2008). On warm conceptual change: The interplay of text, epistemological beliefs, and topic interest. *Journal of Educational Psychology, 100*(2), 291-309.

Masson, S., Potvin, P., Riopel, M., & Foisy, L. M. B. (2014). Differences in brain activation between novices and experts in science during a task involving a common misconception in electricity. *Mind, Brain, and Education, 8*(1), 44-55.

Mateen, F. J., Oh, J., Tergas, A. I., Bhayani, N. H., & Kamdar, B. B. (2013). Titles versus titles and abstracts for initial screening of articles for systematic reviews. *Clinical Epidemiology, 5*, 89-95.

Mathes, E. F., Frieden, I. J., Cho, C. S., & Boscardin, C. K. (2014). Randomized controlled trial of spaced education for pediatric residency education. *Journal of Graduate Medical Education, 6*(2), 270-274.

Matos, J., Petri, C. R., Mukamal, K. J., & Vanka, A. (2017). Spaced education in medical residents: an electronic intervention to improve competency and retention of medical knowledge. *PloS One, 12*(7), e0181418.

Matzie, K. A., Kerfoot, B. P., Hafler, J. P., & Breen, E. M. (2009). Spaced education improves the feedback that surgical residents give to medical students: a randomized trial. *The American Journal of Surgery, 197*(2), 252-257.

Mauelshagen, J., Sherff, C. M., & Carew, T. J. (1998). Differential induction of long-term synaptic facilitation by spaced and massed applications of serotonin at sensory neuron synapses of Aplysia californica. *Learning & Memory, 5*(3), 246-256.

Mavis, B. (2001). Self-efficacy and OSCE performance among second year medical students. *Advances in Health Sciences Education, 6*(2), 93-102.

Maxwell, J. A. (2012). Qualitative research design: An interactive approach (Vol. 41). Newbury Park, CA: Sage publications.

Mayer, R. E. (1975). Information processing variables in learning to solve problems. *Review of Educational Research, 45*(4), 525-541.

Mazur, E. (1997). Peer instruction: Getting students to think in class. In: AIP Conference Proceedings. *Advances in Physiology, 399*(1), 981-988.

McCoubrie, P. (2004). Improving the fairness of multiple-choice questions: a literature review. *Medical Teacher, 26*(8), 709-712.

Menekse, M., Stump, G. S., Krause, S., & Chi, M. T. H. (2013). Differentiated overt learning activities for effective instruction in engineering classrooms. *Journal of Engineering Education, 102*(3), 346-374.

Menzel, R., Manz, G., Menzel, R., & Greggers, U. (2001). Massed and spaced learning in honeybees: the role of CS, US, the intertrial interval, and the test interval. *Learning & Memory, 8*(4), 198-208.

Michael, J. A. (1998). Students' misconceptions about perceived physiological responses. *American Journal of Physiology, 274*(6), S90-S98.

Michael, J.A. (2007). What makes physiology hard for students to learn? Results of a faculty survey. *Advances in Physiology Education, 31*(1), 34-40.

Michael, J. A., Wenderoth, M. P., Modell, H. I., Cliff, W., Horwitz, B., McHale, P., Richardson, D., Silverthorn, D., Williams, S., & Whitescarver, S. (2002). Undergraduates' understanding of cardiovascular phenomena. *Advances in Physiology Education, 26*(2), 72-84.

Miller, K., Lasry, N., Lukoff, B., Schell, J., & Mazur, E. (2014). Conceptual question response times in Peer Instruction classrooms. *Physical Review Special Topics-Physics Education Research, 10*(2), 1-6.

Miller, K., Schell, J., Ho, A., Lukoff, B., & Mazur, E. (2015). Response switching and self-efficacy in Peer Instruction classrooms. *Physical Review Special Topics-Physics Education Research, 11*(1), 1-8.

Miller, J. O., Thammasitboon, S., Hsu, D. C., Shah, M. I., Minard, C. G., & Graf, J. M. (2016). Continuing medical education for air medical providers: the successes and challenges. *Pediatric Emergency Care, 32*(2), 87-92.

Mitchell, E. L., Lee, D. Y., Sevdalis, N., Partsafas, A. W., Landry, G. J., Liem, T. K., & Moneta, G. L. (2011). Evaluation of distributed practice schedules on retention of a newly acquired surgical skill: a randomized trial. *The American Journal of Surgery, 201*(1), 31-39.

Modell, H. I. (2007). Helping students make sense of physiological mechanisms: the "view from the inside". *Advances in Physiology Education, 31*(2), 186-192.

Moore, F. G., & Chalk, C. (2012). Improving the neurological exam skills of medical students. *Canadian Journal of Neurological Sciences, 39*(1), 83-86.

Morgan, P. J., & Cleave-Hogg, D. (2002). Comparison between medical students' experience, confidence and competence. *Medical Education, 36*(6), 534-539.

Morton, J., Anderson, L., Frame, F., Moyes, J., & Cameron, H. (2006). Back to the future: teaching medical students clinical procedures. *Medical Teacher, 28*(8), 723-728.

Moulton, C. A. E., Dubrowski, A., MacRae, H., Graham, B., Grober, E., & Reznick, R. (2006). Teaching surgical skills: what kind of practice makes perfect?: a randomized, controlled trial. *Annals of Surgery, 244*(3), 400-409.

Murad, M. H., Coto-Yglesias, F., Varkey, P., Prokop, L. J., & Murad, A. L. (2010). The effectiveness of self-directed learning in health professions education: a systematic review. *Medical Education, 44*(11), 1057-1068.

Murdoch-Eaton, D., & Whittle, S. (2012). Generic skills in medical education: developing the tools for successful lifelong learning. *Medical Education, 46*(1), 120-128.

Murre, J. M., & Dros, J. (2015). Replication and analysis of Ebbinghaus' forgetting curve. *PloS one, 10*(7), e0120644.

Murrihy, R. C., Byrne, M. K., & Gonsalvez, C. J. (2009). Testing an empirically derived mental health training model featuring small groups, distributed practice and patient discussion. *Medical*

*Education, 43*(2), 140-145.

## N

Nakata, B. N., Cavalini, W., Bonin, E. A., Salvalaggio, P. R., & Loureiro, M. P. (2017). Impact of continuous training through distributed practice for acquisition of minimally invasive surgical skills. *Surgical Endoscopy, 31*(10), 4051-4057.

Naqib, F., Sossin, W. S., & Farah, C. A. (2012). Molecular determinants of the spacing effect. *Neural plasticity*, 1-8.

Nelson, J., & Campbell, C. (2017). Evidence-informed practice in education: meanings and applications. *Educational Research, 59*(2), 127-135.

Nenciovici, L., Brault Foisy, L. M., Allaire-Duquette, G., Potvin, P., Riopel, M., & Masson, S. (2018). Neural Correlates Associated With Novices Correcting Errors in Electricity and Mechanics. *Mind, Brain, and Education, 12*(3), 120-139.

Nesbitt, J. C., St Julien, J., Absi, T. S., Ahmad, R. M., Grogan, E. L., Balaguer, J. M., ... & Putnam, J. B. (2013). Tissue-based coronary surgery simulation: medical student deliberate practice can achieve equivalency to senior surgery residents. *The Journal of Thoracic and Cardiovascular Surgery, 145*(6), 1453-1459.

Nielsen, D. G., Gotzsche, O., Sonne, O., & Eika, B. (2012). The relationship between immediate relevant basic science knowledge and clinical knowledge: Physiology knowledge and transthoracic echocardiography image interpretation. *Advances in Health Sciences Education, 17*(4), 501-513.

Nielsen, K. L., Hansen, G., & Stav, J. B. (2016). How the initial thinking period affects student argumentation during peer instruction: Students' experiences versus observations. *Studies in Higher Education, 41*(1), 124-138.

Nkenke, E., Vairaktaris, E., Bauersachs, A., Eitner, S., Budach, A., Knipfer, C., & Stelzle, F. (2012). Spaced education activates students in a theoretical radiological science course: a pilot study. *BMC Medical Education, 12*(1), 32.

## O

O'Hare, L., Stark, P., McGuinness, C., Biggart, A., & Thurston, A. (2017). Spaced Learning: The Design, Feasibility and Optimisation of SMART Spaces. Evaluation Report and Executive Summary. Education Endowment Foundation.

Ohlsson, S. (2009). Meaning change, multiple routes, and the role of differentiation in conceptual change: alternatives to resubsumption? *Educational Psychologist, 44*(1), 64-71.

Ojha, R., Liu, A., Champion, B. L., Hibbert, E., & Nanan, R. K. H. (2014). Spaced scenario demonstrations improve knowledge and confidence in pediatric acute illness management. *Frontiers in Pediatrics, 2*, 133.

# P

Pagani, M. R., Oishi, K., Gelb, B. D., & Zhong, Y. (2009). The phosphatase SHP2 regulates the spacing effect for long-term memory induction. *Cell, 139*(1), 186-198.

Palizvan, M. R., Nejad, M. R., Jand, A., & Rafeie, M. (2013). Cardiovascular physiology misconceptions and the potential of cardiovascular physiology teaching to alleviate these. *Medical Teacher, 35*(6), 454-458.

Panadero, E. (2017). A review of self-regulated learning: Six models and four directions for research. *Frontiers in Psychology, 8*, 422.

Panzarasa, P., Kujawski, B., Hammond, E. J., & Roberts, C. M. (2016). Temporal patterns and dynamics of e- learning usage in medical education. Educational Technology *Research and Development, 64*(1), 13-35.

Pashler, H., Rohrer, D., Cepeda, N. J., & Carpenter, S. K. (2007). Enhancing learning and retarding forgetting: Choices and consequences. *Psychonomic Bulletin & Review, 14*(2), 187-193.

Patel, V. L., Yoskowitz, N. A., & Arocha, J. F. (2009). Towards effective evaluation and reform in medical education: a cognitive and learning sciences perspective. *Advances in Health Sciences Education, 14*(5), 791-812.

Patel, R., Tarrant, C., Bonas, S., Yates, J., & Sandars, J. (2015). The struggling student: a thematic analysis from the self-regulated learning perspective. *Medical Education, 49*(4), 417-426.

Patocka, C., Khan, F., Dubrovsky, A. S., Brody, D., Bank, I., & Bhanji, F. (2015). Pediatric resuscitation training— Instruction all at once or spaced over time?. *Resuscitation, 88*, 6-11.

Periago, M. C., & Bohigas, X. (2005). A study of second-year engineering students' alternative conceptions about electric potential, current intensity and Ohm's law. *European Journal of Engineering Education, 30*(1), 71-80.

Pernar, L. I., Corso, K., Lipsitz, S. R., & Breen, E. (2013). Using spaced education to teach interns about teaching skills. *The American Journal of Surgery, 206*(1), 120-127.

Pernar, L. I., Beleniski, F., Rosen, H., Lipsitz, S., Hafler, J., & Breen, E. (2012). Spaced education faculty development may not improve faculty teaching performance ratings in a surgery department. *Journal of Surgical Education, 69*(1), 52-57.

Persaud, N., McLeod, P., & Cowey, A. (2007). Post-decision wagering objectively measures awareness. *Nature Neuroscience, 10*(2), 257-261.

Phillips, D. C., Phillips, D. C., & Burbules, N. C. (2000). Postpositivism and educational research. NY: Rowman & Littlefield.

Phillips, J. L., Heneka, N., Bhattarai, P., Fraser, C., & Shaw T. (2019). Effectiveness of the spaced education pedagogy for clinicians' continuing professional development: a systematic review. *Medical Education, 53*(9), 886-902.

Phillips, J. L., Heneka, N., Hickman, L., Lam, L., & Shaw, T. (2017). Can a complex online intervention improve cancer nurses' pain screening and assessment practices? Results from a multicenter, pre-post test pilot study. *Pain Management Nursing, 18*(2), 75-89.

Phillips, J. L., Heneka, N., Hickman, L., Lam, L., & Shaw, T. (2014). Impact of a novel online learning module on specialist palliative care nurses' pain assessment competencies and patients' reports of pain: Results from a quasi-experimental pilot study. *Palliative Medicine, 28*(6), 521-529.

Piaget, J. (1976). Piaget's theory. In: Piaget and His School, edited by Inhelder B, Chipman HH, Zwingmann C. (pp. 11-23). NY: Springer.

Piaget, J. (1978). Success and understanding. Cambridge, MA: Harvard University Press.

Pintrich, P. R. (2002). The role of metacognitive knowledge in learning, teaching, and assessing. *Theory into Practice, 41*(1), 219-225.

Pintrich, P. R., Marx, R. W., & Boyle, R. A. (1993). Beyond cold conceptual change: The role of motivational beliefs and classroom contextual factors in the process of conceptual change. *Review of Educational research, 63*(2), 167-199.

Posner, G. J., Strike, K. A., Hewson, P. W., & Gertzog, W. A. (1982). Accommodation of a scientific conception: Toward a theory of conceptual change. *Science education, 66*(2), 211-227.

Potgieter, M., Malatje, E., Gaigher, E., & Venter, E. (2010). Confidence versus performance as an indicator of the presence of alternative conceptions and inadequate problem-solving skills in mechanics. *International Journal of Science Education, 32*(11), 1407-1429.

Potvin, P. (2013). Proposition for improving the classical models of conceptual change based on neuroeducational evidence: conceptual prevalence. *Neuroeducation, 1*(2), 16-43.

Potvin, P. (2017). The Coexistence Claim and Its Possible Implications for Success in Teaching for Conceptual Change. *European Journal of Science and Mathematics Education, 5*(1), 55-66.

Prince, M. (2004). Does active learning work? A review of the research. *Journal of Engineering Education, 93*(3), 223- 231.

Prince, S. E., Daselaar, S. M., & Cabeza, R. (2005). Neural correlates of relational memory: successful encoding and retrieval of semantic and perceptual associations. *Journal of Neuroscience, 25*(5), 1203-1210.

## Q

-

## R

Raman, M., Mclaughlin, K., Violato, C., Rostom, A., Allard, J. P., & Coderre, S. (2010). Teaching in small portions dispersed over time enhances long-term knowledge retention. *Medical Teacher, 32*(3), 250-255.

Ramani, S. (2006). Twelve tips to promote excellence in medical teaching. *Medical Teacher, 28*(1), 19-23.

Ramsden, P. (2003). Learning to teach in higher education. NY: Routledge.

Rangel, R. H., Miller, L., Sitter, H., Stibane, T., & Strzelczyk, A. (2017). Sure, or unsure? Measuring students' confidence and the potential impact on patient safety in multiple-choice questions. *Medical Teacher, 39*(11), 1189–1194.

Rao, S. P., & DiCarlo, S. E. (2000). Peer instruction improves performance on quizzes. *Advances in Physiology Education, 24*(1), 51–55.

Rapp, E. J., Maximin, S., & Green, D. E. (2014). Practice corner: Retrieval practice makes perfect. *Radiographics, 34*(7), 1869-1870

Raupach, T., Andresen, J. C., Meyer, K., Strobel, L., Koziolek, M., Jung, W., ... & Anders, S. (2016). Test-enhanced learning of clinical reasoning: a crossover randomised trial. *Medical Education, 50*(7), 711-720.

Ravesloot, C.J., van der Schaaf, M.F., Muijtjens, A.M., Haaring, C., Kruitwagen, C. L. J. J., Beek, F. J. A., Bakker, J., van Schaik, J. P., & Ten Cate, T. J. (2015). The don't know option in progress testing. *Advances in Health Sciences Education, 20*(5), 1325-1338.

Rawson, K. A., & Kintsch, W. (2005). Rereading effects depend on time of test. *Journal of Educational Psychology, 97*(1), 70-80.

Razzell, P., & Weinman, J. (1977). The pre-clinical curriculum: an information processing problem. *Medical Education, 11*(4), 257-261.

Regehr, G., & Norman, G. R. (1996). Issues in cognitive psychology: implications for professional education. *Academic Medicine, 71*(9), 988-1001.

Relling, A. E., & Giuliodori, M. J. (2015). Effect of peer instruction on the likelihood for choosing the correct response to a physiology question. *Advances in Physiology Education, 39*(3), 167-171.

Robinson, T., Janssen, A., Kirk, J., DeFazio, A., Goodwin, A., Tucker, K., & Shaw, T. (2017). New approaches to continuing medical education: a QStream (spaced education) program for research translation in ovarian cancer. *Journal of Cancer Education, 32*(3), 476-482.

Roediger III, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences, 15*(1), 20-27.

Rogers, E. M. (2003). Diffusion of Innovations, 5th ed. NY: Simon & Shuster Inc.

Rohrer, D., & Taylor, K. (2007). The shuffling of mathematics problems improves learning. *Instructional Science, 35*(6), 481-498.

Rovick, A. A., Michael, J. A., Modell, H. I., Bruce, D. S., Horwitz, B., Adamson T., Richardson, D. R., Silverthorn, D. U., & Whitescarver, S. A. (1999). How accurate are our assumptions about our students' background knowledge? *The American Journal of Physiology, 276*(6), S93–S101.

Royal College of Physicians and Surgeons of Canada. (2000). CanMEDS 2000: Extract from the CanMEDS 2000 Project Societal Needs Working Group Report. *Medical Teacher, 22*, 549-554.

Rozenshtein, A., Pearson, G. D., Yan, S. X., Liu, A. Z., & Toy, D. (2016). Effect of massed versus interleaved teaching method on performance of students in radiology. *Journal of the American College of Radiology, 13*(8), 979-984.

Ruiter, D. J., van Kesteren, M. T., & Fernandez, G. (2012). How to achieve synergy between medical education and cognitive neuroscience? An exercise on prior knowledge in understanding. *Advances in Health Sciences Education, 17*(2), 225-240.

## S

Sadeh, T., Shohamy, D., Levy, D. R., Reggev, N., & Maril, A. (2011). Cooperation between the hippocampus and the striatum during episodic encoding. *Journal of Cognitive Neuroscience, 23*(7), 1597-1608.

Sandars, J. (2009). The use of reflection in medical education: AMEE Guide. *Medical Teacher, 31*(44), 685-695.

Sandars, J., & Cleary, T. J. (2011). Self-regulation theory: applications to medical education: AMEE Guide No. 58. *Medical Teacher, 33*(11), 875-886.

Sandars, J. (2013). When I say… self-regulated learning. *Medical Education, 47*(12), 1162-1163.

Sandberg, K., Timmermans, B., Overgaard, M., Cleeremans, A. (2010). Measuring consciousness: is one measure better than the other? *Consciousness and Cognition, 19*(4), 1069-1078.

Sayer, R. A. (1992). Method in social science: A realist approach. Newbury Park, CA: Psychology Press.

Scales Jr, C. D., Moin, T., Fink, A., Berry, S. H., Afsar-Manesh, N., Mangione, C. M., & Kerfoot, B. P. (2016). A randomized, controlled trial of team-based competition to increase learner participation in quality-improvement education. *International Journal for Quality in Health Care, 28*(2), 227-232.

Schaverien, M. V. (2010). Development of expertise in surgical training. *Journal of Surgical Education, 67*(1), 37-43.

Schneid, S. D., Pashler H., Armour, C. (2018). How much basic science content do second- year medical students remember from their first year? *Medical Teacher, 41*(2), 231-233.

Schoeff, S., Hernandez, B., Robinson, D. J., Jameson, M. J., & Shonka Jr, D. C. (2017). Microvascular anastomosis simulation using a chicken thigh model: interval versus massed training. *The Laryngoscope, 127*(11), 2490-2494.

Schraw, G., Crippen, K. J., & Hartley, K. (2006). Promoting self-regulation in science education: Metacognition as part of a broader perspective on learning. *Research in Science Education, 36*(1-2), 111-139.

Schraw, G., Dennison, R. S. (1994). Assessing metacognitive awareness. *Contemporary Educational Psychology, 19*(4), 460-475.

Schraw, G., & Moshman, D. (1995). Metacognitive theories. *Educational Psychology Review, 7*(4), 351-371.

Schuwirth, L. W., & van der Vleuten, C. P. (2004). Different written assessment methods: what can be said about their strengths and weaknesses? *Medical Education, 38*(9), 974-979.

Schuwirth, L. W., & van der Vleuten, C. P. (2012). The use of progress testing. *Perspectives on Medical Education, 1*(1), 24-30.

Shaikh, U., Afsar-Manesh, N., Amin, A. N., Clay, B., & Ranji, S. R. (2017). Using an online quiz-based reinforcement system to teach healthcare quality and patient safety and care transitions at the University of California. *International Journal for Quality in Health Care, 29*(5), 735-739.

Sharples, J. (2013). Evidence for the Frontline. London: Alliance for Useful Evidence.

Shaw, T., Long, A., Chopra, S., & Kerfoot, B. P. (2011). Impact on clinical behavior of face-to-face continuing medical education blended with online spaced education: a randomized controlled trial. *Journal of Continuing Education in the Health Professions, 31*(2), 103-108.

Shaw, T. J., Pernar, L. I., Peyre, S. E., Helfrick, J. F., Vogelgesang, K. R., Graydon-Baker, E., ... & Gandhi, T. (2012). Impact of online education on intern behaviour around joint commission national patient safety goals: a randomised trial. *BMJ Quality & Safety, 21*(10), 819-825.

Shea, J. A., Arnold, L., & Mann, K. V. (2004). A RIME perspective on the quality and relevance of current and future medical education research. *Academic Medicine, 79*(10), 931-938.

Shenoi, R., Rubalcava, D., Naik-Mathuria, B., Sloas, H. A., Delemos, D., Xu, L., & Mendez, D. (2016). Interactive spaced online education in pediatric trauma. *Sage Open, 6*(2), 1-8

Shtulman, A., & Lombrozo, T. (2016). Bundles of contradiction: A coexistence view of conceptual change. In D. Barner & A. S. Baron (Ed.), Core knowledge and conceptual change (pp. 53–71) . Oxford, UK: Oxford University Press.

Shtulman, A., & Valcarcel, J. (2012). Scientific knowledge suppresses but does not supplant earlier intuitions. *Cognition, 124*(2), 209-215.

Sigman, M., Pena, M., Goldin, A. P., & Ribeiro, S. (2014). Neuroscience and education: prime time to build the bridge. *Nature Neuroscience, 17*(4), 497-502.

Simanton, E., Hansen, L. (2012). Long-term retention of information across the undergraduate medical school curriculum. *South Dakota Medicine, 65*(7), 261-263.

Simmons, A. L. (2012). Distributed practice and procedural memory consolidation in musicians' skill learning. J*ournal of Research in Music Education, 59*(4), 357-368.

Sinatra, G. M., & Broughton, S. H. (2011). Bridging reading comprehension and conceptual change in science education: The promise of refutation text. *Reading Research Quarterly, 46*(4), 374-393.

Sinitsky, D. M., Fernando, B., & Berlingieri, P. (2012). Establishing a curriculum for the acquisition of laparoscopic psychomotor skills in the virtual reality environment. *The American Journal of Surgery, 204*(3), 367-376.

Smeds, M. R., Thrush, C. R., Mizell, J. S., Berry, K. S., & Bentley, F. R. (2016). Mobile spaced education for surgery rotation improves National Board of Medical Examiners scores. J*ournal of Surgical Research, 201*(1), 99-104.

Smith, M. K., Wood, W. B., Adams, W. K., Wieman, C., Knight, J. K., Guild, N., & Su, T. T. (2009). Why peer discussion improves student performance on in-class concept questions. *Science, 323*(5910), 122-124.

Smith, M. K., Wood, W. B., Krauter, K., & Knight, J. K. (2011). Combining peer discussion with instructor explanation increases student learning from in-class concept questions. *CBE Life Sciences Education, 10*(1), 55- 63.

Smith, R. (2003). Editorial misconduct: Medical editors need effective self regulation. *BMJ, 326,* 1224–1225.

Smith, S. M. (2002). Fast robust automated brain extraction. *Human Brain Mapping, 17*(3), 143-155.

Smith, S. M., Jenkinson, M., Woolrich, M. W., Beckmann, C. F., Behrens, T. E., Johansen-Berg, H., . . . Flitney, D. E. (2004). Advances in functional and structural MR image analysis and implementation as FSL. *Neuroimage, 23*, S208-S219.

Smolen, P., Zhang, Y., & Byrne, J. H. (2016). The right time to learn: mechanisms and optimization of spaced learning. *Nature Reviews Neuroscience, 17*(2), 77-88.

Sodervik, I., Mikkila-Erdmann, M., & Vilppu, H. (2014). Promoting the understanding of photosynthesis among elementary school student teachers through text design. *Journal of Science Teacher Education, 25*(5), 581-600.

Sousa, D. A. (2010). Mind, brain, & education: Neuroscience implications for the classroom. Bloomington, IN: Solution Tree Press.

Sowder, J., Armstrong, B., Lamon, S., Simon, M., Sowder, L., & Thompson, A. (1998). Educating teachers to teach multiplicative structures in the middle grades. *Journal of Mathematics Teacher Education, 1*(2), 127-155.

Spaniol, J., Davidson, P. S., Kim, A. S., Han, H., Moscovitch, M., & Grady, C. L. (2009). Event-related fMRI studies of episodic encoding and retrieval: meta-analyses using activation likelihood estimation. *Neuropsychologia, 47*(8-9), 1765-1779.

Sparck, E. M., Bjork, E. L., & Bjork, R. A. (2016). On the learning benefits of confidence-weighted testing. *Cognitive Research: Principles and Implications, 1*(1), 3.

Spiro, R. J., Vispoel, W. P., Schmitz, J. G., Samarapungavan, A., Boerger, A. E., Britton, B. K., & Glynn, S. M. (1987). Cognitive flexibility and transfer in complex content domains. *Executive control processes in reading*, 177-199.

Spreckelsen, C., & Junger, J. (2017). Repeated testing improves achievement in a blended learning approach for risk competence training of medical students: results of a randomized controlled trial. *BMC Medical Education, 17*(1), 177.

Spruit, E. N., Band, G. P., & Hamming, J. F. (2015). Increasing efficiency of surgical training: effects of spacing practice on skill acquisition and retention in laparoscopy training. *Surgical Endoscopy, 29*(8), 2235-2243.

Sreenivasulu, B., & Subramaniam, R. (2014). Exploring undergraduates' understanding of transition metals chemistry with the use of cognitive and confidence measures. *Research in Science Education, 44*(6), 801-828.

Stahl, S. M., Davis, R. L., Kim, D. H., Lowe, N. G., Carlson, R. E., Fountain, K., & Grady, M. M. (2010). Play it again: The master psychopharmacology program as an example of interval learning in bite-sized portions. *CNS Spectrums, 15*(8), 491-504.

Stefanidis, D., Korndorffer Jr, J. R., Markley, S., Sierra, R., & Scott, D. J. (2006). Proficiency maintenance: impact of ongoing simulator training on laparoscopic skill retention. *Journal of the American College of Surgeons, 202*(4), 599-603.

Stefanidis, D., Walters, K. C., Mostafavi, A., & Heniford, B. T. (2009). What is the ideal interval between training sessions during proficiency-based laparoscopic simulator training?. *The American Journal of Surgery, 197*(1), 126- 129.

Stransky, D., Wilcox, L. M., & Dubrowski, A. (2010). Mental rotation: Cross-task training and generalization. *Journal of Experimental Psychology: Applied, 16*(4), 349-360.

Sullivan, P. B., Gregg, N., Adams, E., Rodgers, C., Hull, J. (2013). How much of the paediatric core curriculum do medical students remember? *Advances in Health Sciences Education, 18*(3), 365-373.

Swartz, S. M. (2006). Acceptance and Accuracy of Multiple Choice, Confidence-Level, and Essay Question Formats for Graduate Students. *Journal of Education for Business, 81*(4), 215-220.

# T

Tanner K. D., & Allen D. (2005). Approaches to biology teaching and learning: understanding the wrong answers--teaching toward conceptual change. *CBE Life Sciences Education, 4*(2), 112-117.

Tanner, K. D. (2012). Promoting student metacognition. *CBE Life Sciences Education, 11*(2), 113-120.

Taveira-Gomes, T., Saffarzadeh, A., Severo, M., Guimarães, M. J., & Ferreira, M. A. (2014). A novel collaborative e- learning platform for medical students-ALERT STUDENT. *BMC Medical Education, 14*(1), 143.

Taveira-Gomes, T., Prado-Costa, R., Severo, M., & Ferreira, M. A. (2015). Characterization of medical students recall of factual knowledge using learning objects and repeated testing in a novel e-learning system. *BMC medical education, 15*(1), 4.

Taylor, A. K., & Kowalski, P. (2004). Naive psychological science: the prevalence, strength, and sources of misconceptions. *Psychological Record, 54*(1), 15–25.

Terenyi, J., Anksorus, H., & Persky, A. M. (2018). Impact of spacing of practice on learning brand name and generic drugs. *American Journal of Pharmaceutical Education, 82*(1), 6179.

Thiede, K. W., Anderson, M., & Therriault, D. (2003). Accuracy of metacognitive monitoring affects learning of texts. *Journal of Educational Psychology, 95*(1), 66-73.

Thiede, K. W., Griffin, T. D., Wiley, J., & Anderson, M. C. (2010). Poor metacomprehension accuracy as a result of inappropriate cue use. *Discourse Processes, 47*(4), 331-362.

Thiede, K. W., Griffin, T. D., Wiley, J., & Redford, J. S. (2009). Metacognitive monitoring during and after reading. In: Handbook of metacognition in education. (pp. 97-118). NY: Routledge.

Thomas, M. S. (2019). Response to Dougherty and Robey (2018) on Neuroscience and Education: Enough Bridge Metaphors—Interdisciplinary Research Offers the Best Hope for Progress. *Current Directions in Psychological Science, 28*, 337-340.

Thomas, A., Lubarsky, S., Durning, S. J., & Young, M. E. (2017). Knowledge syntheses in medical education: demystifying scoping reviews. *Academic Medicine, 92*(2), 161-166.

Thomas, A., D. Gruppen, L., van der Vleuten, C., Chilingaryan, G., Amari, F., & Steinert, Y. (2019a).

Use of evidence in health professions education: Attitudes, practices, barriers and supports. *Medical Teacher, 41*(9), 1012-1022.

Thomas, M. S., Ansari, D., & Knowland, V. C. (2019b). Annual Research Review: Educational neuroscience: progress and prospects. *Journal of Child Psychology and Psychiatry and Allied Disciplines, 60*(4), 477-492.

Thorndike, E. (1926). Educational psychology. Volume 1: The original nature of man. In: Teachers College, New York.

Tippett, C. D. (2010). Refutation text in science education: A review of two decades of research. *International Journal of Science and Mathematics Education, 8*(6), 951-970.

Tommerdahl, J. (2010). A model for bridging the gap between neuroscience and education. *Oxford Review of Education, 36*(1), 97-109.

Treagust, D. F. (1986). Evaluating students' misconceptions by means of diagnostic multiple choice items. *Research in Science Education, 16*(1), 199-207.

Treagust, D. F. (1988). Development and use of diagnostic tests to evaluate students' misconceptions in science. *International Journal of Science Education, 10*(2), 159-169.

Trotskovsky, E., Waks, S., Sabag, N., & Hazzan, O. (2013). Students' misunderstandings and misconceptions in engineering thinking. *International Journal of Engineering Education, 29*(1), 107-118.

Tse, D., Langston, R.F., Kakeyama, M., Bethus, I., Spooner, P. A., Wood, E. R., Witter, M. P., & Morris, R. G. (2007). Schemas and memory consolidation. *Science, 316*(5821), 76-82.

Tshibwabwa, E., Mallin, R., Fraser, M., Tshibwabwa, M., Sanii, R., Rice, J., & Cannon, J. (2017). An Integrated Interactive-Spaced Education Radiology Curriculum for Preclinical Students. *Journal of Clinical Imaging Science*, *7*, 22.

Tsui, C. Y., & Treagust, D. F. (2010). Evaluating secondary students' scientific reasoning in genetics using a two-tier diagnostic instrument. *International Journal of Science Education, 32*(8), 1073-1098.

Tweed, M. J., Purdie, G., & Wilkinson, T. (2017a). Low performing students have insightfulness when they reflect- in-action. *Medical Education, 51*(3), 316-323.

Tweed, M. J., Stein, S., Wilkinson, T. J., Purdie, G., & Smith, J. (2017b). Certainty and safe consequence responses provide additional information from multiple choice question assessments. *BMC Medical Education, 17*(1), 106.

## U

-

## V

Vaughn, A. R., Brown, R. D., & Johnson, M. L. (2020). Understanding Conceptual Change and Science Learning through Educational Neuroscience. *Mind, Brain, and Education, 14*(2), 82-93.

Verdaasdonk, E. G. G., Stassen, L. P. S., Van Wijk, R. P. J., & Dankelman, J. (2007). The influence of different training schedules on the learning of psychomotor skills for endoscopic surgery. *Surgical Endoscopy, 21*(2), 214- 219.

Vergnaud, G. (1983). Multiplicative structures. In R. Lesh & M. Landau (Eds.), Acquisition of mathematics concepts and processes (pp. 127-174). Orlando, FL: Academic Press, Inc.

Versteeg, M., van Blankenstein, F. M., Putter, H., & Steendijk, P. (2019). Peer instruction improves comprehension and transfer of physiological concepts: a randomized comparison with self-explanation. *Advances in Health Sciences Education, 24*(1), 151-165.

Versteeg, M., Wijnen-Meijer, M., Steendijk, P. (2019). Informing the uninformed: a multitier approach to uncover students' misconceptions on cardiovascular physiology. *Advances in Physiology Education, 43*(1), 7-14.

Vickrey, T., Rosploch, K., Rahmanian, R., Pilarz, M., & Stains, M. (2015). Research-based implementation of peer instruction: A literature review. *CBE Life Sciences Education, 14*(1), 11.

van der Vleuten, C. P. M., Dolmans, D. H. J. M, Scherpbier, A. J. J. A. (2000). The need for evidence in education. *Medical Teacher, 22*(3), 246-250.

Vosniadou, S. (1994). Capturing and modeling the process of conceptual change. *Learning and Instruction, 4*(1), 45-69.

## W

Wandersee, J. H., Mintzes, J. J., & Novak, J. D. (1994). Research on alternative conceptions in science. In: Handbook Of Research On Science Teaching And Learning, edited by Gabel DL. (pp. 177–210). NY: Macmillan.

Weggemans, M. M., Custers, E. J., ten Cate, O. T. J. (2017). Unprepared retesting of first year knowledge: how much do second year medical students remember? *Medical Science Educator, 27*(4), 597–605.

Weidman, J., & Baker, K. (2015). The cognitive science of learning: concepts and strategies for the educator and learner. *Anesthesia & Analgesia, 121*(6), 1586-1599.

Weinstein, Y., Madan, C. R., & Sumeracki, M. A. (2018). Teaching the science of learning. *Cognitive Research: Principles and Implications, 3*(1), 2-17.

Weisberg, D. S., Keil, F. C., Goodstein, J., Rawson, E. & Gray, J. R. (2008). The seductive allure of neuroscience explanations. *Journal of Cognitive Neuroscience, 20*(3), 470-477.

Weisberg, D. S., Taylor, J. C. & Hopkins, E. J. (2015). Deconstructing the seductive allure of neuroscience explanations. *Judgment and Decision Making 10*(5), 429-441.

Wierzchon, M., Wronka, E., Paulewicz, B., & Szczepanowski, R. (2016). Post-decision wagering affects metacognitive awareness of emotional stimuli: an event related potential study. *PloS One, 11*(8), e159516.

Wiltshire, G. (2018). A case for critical realism in the pursuit of interdisciplinarity and impact. *Qualitative Research in Sport, Exercise and Health 10*, 525-542.

Wong, R. M., Lawson, M. J., & Keeves, J. (2002). The effects of self-explanation training on students' problem solving in high-school mathematics. *Learning and Instruction, 12*(2), 233-262.

Wood, A. K., Galloway, R. K., Hardy, J., & Sinclair, C. M. (2014). Analyzing learning during Peer Instruction dialogues: A resource activation framework. *Physical Review Special Topics-Physics Education Research, 10*(2), 020107.

Woods, N. N. (2007). Science is fundamental: The role of biomedical knowledge in clinical reasoning. *Medical Education, 41*(12), 1173-1177.

Woods, N. N., Brooks, L. R., & Norman, G. R. (2007). It all make sense: Biomedical knowledge, causal connections and memory in the novice diagnostician. *Advances in Health Sciences Education, 12*(4), 405-415.

Woolrich, M. W., Behrens, T. E., Beckmann, C. F., Jenkinson, M., & Smith, S. M. (2004). Multilevel linear modelling for FMRI group analysis using Bayesian inference. *Neuroimage, 21*(4), 1732-1747.

Wormald, B. W., Schoeman, S., Somasunderam, A., & Penn, M. (2009). Assessment drives learning: an unavoidable truth?. *Anatomical Sciences Education, 2*(5), 199-204.

## X

Xie, C., Bai, F., Yu, H., Shi, Y., Yuan, Y., Chen, G., . . . Li, S.-J. (2012). Abnormal insula functional network is associated with episodic memory decline in amnestic mild cognitive impairment. *Neuroimage, 63*(1), 320-327.

Xue, G., Mei, L., Chen, C., Lu, Z. L., Poldrack, R., & Dong, Q. (2011). Spaced learning enhances subsequent recognition memory by reducing neural repetition suppression. *Journal of Cognitive Neuroscience, 23*(7), 1624- 1633.

## Y

Yan, Y. K., & Subramaniam, R. (2018). Using a multi-tier diagnostic test to explore the nature of students' alternative conceptions on reaction kinetics. *Chemistry Education Research and Practice, 19*(1), 213-226.

Yeh, D.D., Park, Y.S. (2015). Improving learning efficiency of factual knowledge in medical education. *Journal of Surgical Education, 72*(5), 882-889.

Young, M., Thomas, A., Lubarsky, S., Ballard, T., Gordon, D., Gruppen, L. D., ... & Durning, S. J. (2018). Drawing boundaries: the difficulty in defining clinical reasoning. *Academic Medicine, 93*(7), 990-995.

Young, J. Q., Van Merrienboer, J., Durning, S., & Ten Cate, O. (2014). Cognitive load theory: implications for medical education: AMEE Guide No. 86. *Medical Teacher, 36*(5), 371-384.

Ystad, M., Eichele, T., Lundervold, A. J., & Lundervold, A. (2010). Subcortical functional connectivity and verbal episodic memory in healthy elderly—a resting state fMRI study. *Neuroimage, 52*(1), 379-388.

## Z

Zhang, N., & Henderson, C. N. R. (2016). Brief, cooperative peer-instruction sessions during lectures enhance student recall and comprehension. *Journal of Chiropractic Education, 30*(2), 87-93.

Ziegler, B., & Montplaisir, L. (2014). Student Perceived and Determined knowledge of Biology Concepts in an Upper-Level Biology Course. *CBE Life Sciences Education, 13*(2), 322-330.

Zimmerman, B. J. (1990). Self-regulated learning and academic achievement: An overview. *Educational Psychologist, 25*(1), 3-17.

Zimmerman, B. J. (2000). Attaining self-regulation: A social cognitive perspective. In: Handbook of self-regulation (pp. 13-39). San Diego, CA: Academic Press.

Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. *Theory Into Practice, 41*(2), 64-70.

Zimmerman, B. J., & Schunk, D. H. (2013). Reflections on theories of self-regulated learning and academic achievement. In: Self-regulated learning and academic achievement (pp. 282-301). NY: Routledge.

Zohar, A., & Barzilai, S. (2013). A review of research on metacognition in science education: Current and future directions. *Studies in Science Education, 49*(2), 121-169.

# List of scientific contributions

**Scientific articles**

Versteeg, M., Hendriks, R. A., Thomas, A., Ommering, B. W. C., & Steendijk, P. (2020). Conceptualising spaced learning in health professions education: A scoping review. *Medical Education, 54*(3), 205-216.

Versteeg, M., & Steendijk, P. (2019). Putting post-decision wagering to the test: a measure of self-perceived knowledge in basic sciences? *Perspectives on Medical Education, 8*(1), 9-16.

Versteeg, M., van Blankenstein, F. M., Putter, H., & Steendijk, P. (2019). Peer instruction improves comprehension and transfer of physiological concepts: a randomized comparison with self-explanation. *Advances in Health Sciences Education, 24*(1), 151-165.

Versteeg, M., Wijnen-Meijer, M., & Steendijk, P. (2019). Informing the uninformed: a multitier approach to uncover students' misconceptions on cardiovascular physiology. *Advances in physiology education, 43*(1), 7-14.

Timmer, M. C. J., Steendijk, P., Arend, S. M., & Versteeg, M. (2020). Making a lecture stick: the effect of spaced instruction on knowledge retention in medical education. *Medical Science Educator*, *30,* 1211-1219.

Versteeg M., van Loon M., Wijnen-Meijer M., & Steendijk P. Refuting misconceptions in medical physiology. *BMC Medical Education*, *20*(1), 1-9.

Versteeg M., Guusje B., Wijnen-Meijer M., Ommering B. W. C., de Beaufort A. J., Steendijk P. What were you thinking? Medical students' metacognition and perceptions of self-regulated learning. *Submitted.*

Versteeg M., & Steendijk P. The origins of medical students' misconceptions and misunderstandings in cardiovascular physiology. *Submitted*.

Versteeg M., Hafkemeijer A., de Beaufort A. J., & Steendijk P. An understanding of (mis)understanders: exploring the underlying mechanisms of concept learning using functional magnetic resonance imaging. *Submitted.*

## Awards & nominations

| | | |
|---|---|---|
| 2019 | Spaced learning in het medisch onderwijs: een scoping review. | NVMO congress (nominated) |
| 2018 | Door de ogen van de student: misconcepties in het medisch fysiologisch onderwijs | NVMO congress (paper prize) |
| 2018 | Spaced learning in het medisch onderwijs: tijd voor een instructionele revolutie? | NVMO congress (poster prize) |
| 2017 | Actieve leerstrategieën voor toepassingen van fysiologische concepten in medisch onderwijs | NVMO congress (nominated) |

## Presentations

| | | |
|---|---|---|
| 2020 | Spaced learning in medical education: a scoping review | Cardiology research reports |
| 2020 | An understanding of misunderstanders | OEC research meeting |
| 2019 | Leren slim te leren. Meer aandacht voor metacognitie in medisch onderwijs | NVMO congress |
| 2019 | Spaced learning in het medisch onderwijs: een scoping review | NVMO congress |
| 2019 | Making a lecture stick: The effects of spaced instruction on knowledge retention in medical education | AMEE congress |
| 2019 | Informing the uninformed: students' misconceptions in cardiovascular physiology | Cardiology research reports |
| 2018 | Door de ogen van de student: misconcepties in het medisch fysiologisch onderwijs | NVMO congress |
| 2018 | Weet jij wat je (niet) weet? De Post-Decision Wagering methode als een maat voor zelfkennis | NVMO congress |
| 2018 | Through the lens of confidence: students' misconceptions in physiology | EARLI SIG 03 meeting |
| 2018 | Hoe ervaren geneeskunde studenten hun leeruitkomsten van de studie? | NVMO PhD day |
| 2018 | A typology of reviews | OEC research meeting |
| 2018 | Active learning strategies for comprehension and transfer of physiological concepts | Cardiology research reports |
| 2017 | Actieve leerstrategieën voor toepassingen van fysiologische concepten in medisch onderwijs | NVMO congress |
| 2017 | Evidence-based education in medical physiology | Rogano meeting |
| 2017 | An experimental approach to medical education | NVMO PhD day |
| 2017 | Contexts, concepts and cognition | OEC research meeting |

## Posters

| | | |
|---|---|---|
| 2018 | Spaced learning in het medisch onderwijs | NVMO congress |
| 2018 | Moving towards spaced learning in medical education: time to start an instructional revolution? | EARLI SIG 22 meeting |
| 2016 | Toepassing van 'peer discussion' tijdens werkgroep onderwijs | NVMO congress |

## (Invited) lectures

| | | |
|---|---|---|
| 2020 | Through the lens of confidence: students' misconceptions in physiology | Technical University of Munich & Leiden Institute of Advanced Computer Science |
| 2020 | Onderzoek van medisch onderwijs | LUMC Honours programme |
| 2020 | Leren slim te leren: meer aandacht voor metacognitie in medisch onderwijs | LUMC LEARN |
| 2019 | From bench to classroom: investigating concept learning in medical education | LUMC Groot Onderwijsoverleg |
| 2019 | Making a lecture stick: The effects of spaced instruction on knowledge retention in medical education | James Pickering visit |
| 2019 | Learning to learn | Law Faculty |
| 2019 | Onderzoek van onderwijs | Student Advies Commissie |
| 2019 | Zelfreguleren kun je leren: een (neuro) wetenschappelijke kijk op metacognitie | LUMC LEARN |
| 2019 | Education meets Neuroscience: implications for teaching and learning | Law Faculty |
| 2019 | Onderzoek van onderwijs | Board of Directors |
| 2018 | Misconcepties in het medisch fysiologisch onderwijs | LUMC programme committee |
| 2018 | Education meets Neuroscience: implications for teaching and learning | Leiden Faculty of Science & LUMC LEARN |
| 2018 | Insights on how neuroscience can change your teaching | Leiden Teachers Academy |
| 2017 | Simulations in Healthcare | Clinical Technology Bachelor |
| 2017 | Does 'Peer Instruction' enhance understanding and transfer of physiological concepts? | LUMC LEARN |

# Dankwoord

"Ik heb het nog nooit gedaan, dus ik denk dat ik het wel kan." En ja, het is gelukt! Dat ik de eindstreep heb gehaald is mede te danken aan fijne collega's, vrienden en familie. Een aantal van deze mensen wil ik graag in het bijzonder bedanken door ze te noemen op deze bladzijden in dit proefschrift, die waarschijnlijk als eerste gelezen zullen worden.

Prof.dr. Steendijk, beste **Paul**, jij was voor mij alles-in-1. Naast een toegewijde promotor, was jij mijn enthousiaste mede-onderzoeker en betrouwbare 'dagelijkse' begeleider. Veel dank daarvoor. Ik kreeg van jou de vrijheid om mijn eigen koers te bepalen. Dat vertrouwen leverde een prettige samenwerking op, waar ik met een fijn gevoel op terug kijk.

Dr. de Beaufort, beste **AJ**, bedankt voor je begeleiding als co-promotor tijdens mijn eindsprint. Jouw enthousiasme werkte voor mij als een uitstekende motivator!

Co-auteurs, **AJ**, **Anne**, **Aliki**, **Belinda**, **Floris, Guusje**, **Hein**, **Mariëtte**, **Marjo**, **Marnix**, **Paul**, **Renée** en **Sandra**, dank voor het delen van jullie expertise en het mede vormgeven van de publicaties in dit proefschrift.

Collega's van de CRIME onderzoeksgroep, bedankt voor jullie betrokkenheid bij het onderwijskundig onderzoek. Ik voel me trots als ik zie dat we samen een nieuwe wetenschappelijke discipline binnen het LUMC op de kaart hebben gezet. **Marchien**, jij hebt mijn inziens aangetoond dat een manager voor een goed functionerende onderzoeksgroep van onschatbare waarde is. Dank voor al je inspanningen en de inspirerende gesprekken.

Collega's van het Onderwijs Expertise Centrum, bedankt voor alle onderwijskundige kennis en inspiratie die ik bij jullie heb opgedaan. Ik kijk met plezier terug op de werkoverleggen, Journal Clubs, praatjes bij het tostiapparaat, uitjes en (kerst) borrels. Ik zal de OECmagie zeker niet snel vergeten! **Lars**, heel erg bedankt voor je hulp bij de layout van mijn proefschrift, zonder jou had deze er een heel stuk saaier uitgezien!

Collega's van de afdeling Hartziekten, dank voor het door jullie in mij gestelde vertrouwen om namens de Hartziekten te pionieren in het onderwijskundig onderzoek in het LUMC.

Collega's van de afdeling Radiologie en Faculteit Sociale Wetenschappen, **Annette, Jordi, Marissa, Mischa, Philip, Wouter** en alle **student-assistenten**, ontzettend bedankt voor alle tijd en energie die jullie hebben gestoken in die verdwaalde vreemdeling met haar out-of-the-box onderwijsexperiment. In het bijzonder gaat mijn dank uit naar **Anne**, dank dat jouw welwillendheid om mij te adopteren als verloren MRI-onderzoeker-in-spe is uitgegroeid tot een succesvolle samenwerking en een mooie vriendschap.

Onderwijskundig onderzoekers van de PHEG, dank voor het delen van jullie expertise tijdens onze bijeenkomsten. In het bijzonder **Irene**, ik heb je leren kennen als een wijze vrouw die voor mij klaarstond met goede raad. Een onderzoeker om van te blijven leren!

JC'ers, bedankt voor alle discussies die we de afgelopen jaren hebben gevoerd over ons onderzoek. Ik wil jullie ook graag bedanken voor alle borrels, etentjes en gezelligheid! **Jan-Willem**, voor jou een persoonlijk dankjewel dat jij mijn medevoorzitter wilde zijn, de overleggen met de Lebkov koffietjes zijn fijne herinneringen.

Leden van de NVMO promovendidagcommissie, **Ellen, Iris, Jolien, Joyce, Lubberta en Margot**, dank dat jullie mijn commissiegenootjes waren. De afspraakjes op centraal station (en daarbuiten) waren altijd gezellige momenten om even bij te kletsen over onze promotieperikelen. **Iris**, jou bedank ik graag ook voor je rol als medevoorzitter van het promovendinetwerk. Wat was het fijn om dat samen met jou op te pakken!

De allerleukste feestcommissiegenootjes, **Belinda, Harmke, Petra en Renée,** ik zou graag voor altijd met een gouden glitterjasje over de V7 lopen om het feestgevoel te verspreiden. Grapje natuurlijk, maar ik mis onze HUF attributen wel. Dank voor alle momenten van ontspanning.

Mijn dierbare mede-promovendi van de V7-38, **Belinda, Charlotte, Katja, Kirsten, Sandhia en Renée**, Onze open werksfeer is een van die dingen die mijn tijd in het LUMC zo aangenaam heeft gemaakt. Bedankt voor alle onvergetelijke momenten!

**Charlotte en Renée,** mijn paranimfen, wat fijn dat jullie op de grote dag achter mij willen staan. Jullie enthousiasme voor onderzoekparadigma's werkte de afgelopen jaren erg aanstekelijk en onze discussies daarover geven mij extra vertrouwen bij mijn verdediging en in mijn verdere carrière als onderzoeker!

Lieve **Matthijs**, bedankt voor alle liefde die jij mij geeft. Ik voel me erg gelukkig met jou en kijk uit naar wat de toekomst ons brengt!

Lieve **Papa** en **Mama**, bedankt voor jullie onvoorwaardelijke steun. Ik waardeer het erg dat jullie zoveel voor mij doen en dat ik altijd op jullie kan rekenen. **Pim**, mijn broertje, jouw doorzettingsvermogen is bewonderenswaardig, ik kijk uit naar alle avonturen die jij nog gaat beleven.

# Curriculum Vitae

Marjolein Versteeg was born on the 12th of October 1994 in Hengelo, the Netherlands. After completing high school at Bataafs Lyceum in 2011, she received her bachelor's degree in Biomedical Sciences from Utrecht University in 2014. Because of her interest in neurobiology, Marjolein enrolled in the research master Neuroscience & Cognition. In 2016, she received her master's degree from Utrecht University. As part of her master's degree, Marjolein performed molecular biology research regarding neurodegenerative diseases at the Utrecht University Medical Center and Harvard University. Upon her return from the United States, Marjolein started her PhD research at Leiden University Medical Center under supervision of Prof.dr. Paul Steendijk. In her PhD project, Marjolein studied various learning processes from an educational neuroscience point of view. Over the years, she developed into a medical education researcher with knowledge about a broad range of educational research methods and methodologies. Her findings resulted in more insights in learning and instruction within the context of health professions education as described in this thesis. Besides her research activities, Marjolein was actively involved in the PhD network of the Dutch Association for Medical Education. In September 2020, Marjolein started a new job at VeiligheidNL, where she will continue working as a researcher.