Accuracy of Self-Evaluation in Biology Education: Investigating Potential Effects of Skill Practice, Prompting, and Physical Exercise

Doctoral Thesis

for

the award of the doctoral degree

of the Faculty of Mathematics and Natural Sciences

of the University of Cologne

submitted by

Stefanie Elsner

in the year 2025, Cologne

Reviewers:

Prof. Dr. Jörg Großschedl

Prof. Dr. Benjamin Rott

Knowing yourself is the beginning of all wisdom.

Aristotle

List of Contents

Abstract

Enabling students to become life-long learners is an overarching educational goal. Self-evaluations are an integral part of self-regulated learning, a key concept used to support life-long learning. The accuracy of self-evaluation plays a major role in learning processes, as it can influence subsequent learning behaviour. Yet, inaccurate self-evaluations are frequently observed across disciplines and contexts, including biology education. Deliberate, theory-and evidence-based approaches are needed to enhance the accuracy of self-evaluation. This dissertation comprises two published studies and one unpublished manuscript that examine three interdisciplinary approaches potentially affecting the accuracy of self-evaluation. Study I examined the potential effects of skill practice on the accuracy of self-evaluation. In this controlled, quasi-experimental intervention study, 167 eighth-grade students took part in a concept map construction, concept map study, or control training. They evaluated their concept mapping skills in a subsequent cross-over learning phase. The accuracy of self-evaluation was measured using correlations. The study findings revealed a slightly increased alignment of self-evaluation and “objective” concept mapping skills after concept map construction training compared to concept map study training. Study II examined the potential effects of prompting on the accuracy of self-evaluation. In this controlled and randomised online study, 162 pre-service teachers were prompted to use resource-oriented and deficit-oriented comprehension questions while reading a biology text. They evaluated their text comprehension before taking a short comprehension test. The accuracy of self-evaluation was determined by calculating the difference between subjective self-evaluation and “objective” performance. The results showed no evidence for an effect of prompting with resource-oriented or deficit-oriented self-questions. Study III examined the potential acute effects of physical exercise on the accuracy of self-evaluation. In this controlled, within-subjects laboratory study, 24 children with Attention Deficit Hyperactivity Disorder (ADHD) took part in strength-based, flexibility-based, and control training. They evaluated their task accuracy and reaction time in a subsequent Eriksen Flanker Task, which measured attentional performance. The accuracy of self-evaluation of task accuracy was measured by calculating the difference between subjective self-evaluation and “objective” task accuracy. The accuracy of self-evaluation of reaction time was determined using a categorisation of correct and incorrect self-evaluations. The findings of Study III showed no evidence for altered attentional performance or positive effect on the accuracy of self-evaluation. The results of the studies included in this dissertation support previous findings indicating an increase in the accuracy of self-evaluation after practicing the skill to be evaluated. To contextualise the results, the methodological approaches are critically discussed, and the role of non-significant research findings is addressed. This dissertation provides interdisciplinary insights into the complexity of self-evaluation and self-regulated learning by applying approaches from cognitive and educational psychology to biology education.

Zusammenfassung

Ein übergreifendes Bildungsziel ist es, Lernende zum lebenslangen Lernen zu befähigen. Selbsteinschätzungen sind ein wesentlicher Bestandteil des selbstregulierten Lernens, ein Schlüsselkonzept zur Förderung des lebenslangen Lernens. Die Genauigkeit der Selbsteinschätzungen spielt eine bedeutsame Rolle für Lernprozesse, auch weil sie das anschließende Lernverhalten beeinflussen kann. Dennoch werden häufig ungenaue Selbsteinschätzungen in verschiedenen Disziplinen und Kontexten beobachtet, ebenso im Biologieunterricht. Gezielte theorie- und evidenzbasierte Ansätze sind notwendig, um die Genauigkeit von Selbsteinschätzungen zu fördern. Diese Dissertation umfasst zwei veröffentlichte Studien und ein unveröffentlichtes Manuskript, die drei interdisziplinäre Ansätze auf ihre potenziellen Effekte hinsichtlich der Genauigkeit von Selbsteinschätzungen untersuchen. Studie I untersuchte die potenziellen Effekte des praktischen Übens einer Fähigkeit auf die Genauigkeit der Selbsteinschätzung. In einer kontrollierten, quasi-experimentellen Interventionsstudie nahmen 167 Schüler:innen der achten Klasse entweder an einem Training zur Konstruktion von Concept Maps, einem Training zur Betrachtung von Concept Maps oder einem Kontrolltraining teil. Die Schüler:innen beurteilten in einer anschließenden Cross-over-Lernphase ihre Fähigkeiten im Concept Mapping. Die Genauigkeit der Selbsteinschätzung wurde anhand von Korrelationsanalysen bestimmt. Die Ergebnisse der Studie zeigen eine leicht erhöhte Übereinstimmung zwischen Selbsteinschätzung und „objektiver“ Messung nach dem Konstruktionstraining im Vergleich zum Training, in dem die Betrachtung von Concept Maps geübt wurde. Studie II untersuchte die potenziellen Effekte des Promptings auf die Genauigkeit der Selbsteinschätzung. In einer kontrollierten und randomisierten Online-Studie mit 162 Studierenden des Lehramts wurden während des Lesens eines Biologielehrbuchtextes Prompts verwendet. Diese Prompts wiesen darauf hin entweder eine ressourcenorientierte, eine defizitorientierte oder keine Frage an sich selbst zu richten und zu beantworten. Die Studierenden beurteilten ihr Textverständnis und bearbeiteten einen kurzen Test zum Leseverständnis. Die Genauigkeit der Selbsteinschätzung wurde anhand der Differenz zwischen subjektiver Selbsteinschätzung und „objektiver“ Leistung bestimmt. Die Ergebnisse zeigen keine Evidenz für einen Effekt des Promptings mit ressourcenorientierten und defizitorientierten Fragen an sich selbst. Studie III untersuchte akute Effekte sportlicher Aktivität auf die Genauigkeit von Selbsteinschätzungen. In einer kontrollierten Laborstudie mit Messwiederholung nahmen 24 Kinder mit einer Diagnose der

Aufmerksamkeitsdefizit-Hyperaktivitätsstörung (ADHS) an einem kraftbasierten Training, einem flexibilitätsbasierten Training und einem Kontrolltraining teil. Sie beurteilten die Aufgabengenauigkeit und die Reaktionszeit in einer daran anschließenden Eriksen Flanker Task, einer Aufgabe zur Messung der Aufmerksamkeitsperformanz. Die Genauigkeit der Selbsteinschätzung der Aufgabengenauigkeit wurde mithilfe der Differenz zwischen subjektiver Selbsteinschätzung und „objektiver“ Aufgabengenauigkeit bestimmt. Die Genauigkeit der Selbsteinschätzung der Reaktionszeit wurde anhand einer Kategorisierung korrekter und inkorrekter Selbsteinschätzungen ermittelt. Die Ergebnisse zeigten weder Evidenz für eine veränderte Aufmerksamkeitsperformanz noch für einen positiven Effekt auf die Genauigkeit der Selbsteinschätzung. Die Ergebnisse der Studien dieser Dissertation unterstützen zuvor veröffentlichte wissenschaftliche Befunde, die eine Verbesserung der Selbsteinschätzung nach dem Üben der einzuschätzenden Fähigkeit zeigen. Um die Ergebnisse einzuordnen, werden die methodischen Herangehensweisen kritisch diskutiert, und die Bedeutsamkeit von nicht-signifikanten Forschungsergebnissen adressiert. Diese Dissertation stellt eine interdisziplinäre Sichtweise auf die Komplexität von Selbsteinschätzungen bereit, indem sie kognitions- und pädagogisch-psychologische Erklärungsansätze auf die Biologiedidaktik anwendet.

Introduction

"Knowing yourself is the beginning of all wisdom."

Aristotle

Undoubtedly, knowing ourselves has value in itself. Moreover, knowing ourselves can be a powerful tool for learning. Understanding how we learn and judging our own learning process accurately determine – at least partially – our future learning behaviour (Metcalfe & Finn, 2008). Self-evaluation is defined as "the act or process of judging your own abilities and performance" (Cambridge Advanced Learner's Dictionary & Thesaurus, 2024). In this dissertation, the accuracy of self-evaluation is understood as the congruence between subjective evaluations of one's own learning and "objectively" measured learning parameters.[1] The ability to accurately self-evaluate is neither innate nor self-evident. Inaccurate self-evaluations are ubiquitous as shown by various phenomena: The Dunning-Kruger effect, the positive illusory bias, and the big-fish-little-pond effect exemplify over- and underestimation in diverse contexts (Dunning, 2011; Marsh, 1987; Owens et al., 2007). The Dunning-Kruger effect, for instance, shows that people scoring in the bottom quartile (12th percentile) across different disciplines such as humour, grammar, and logic rate their skills as above average (62nd percentile; Kruger & Dunning, 1999). The positive illusory bias (PIB) describes overly positive self-evaluations by children with ADHD compared to observations by their parents or teachers, particularly in areas where the children show impairments (Hoza et al., 2004; Hoza et al., 2002). The big-fish-little-pond effect (BFLPE) suggests that students misjudge their abilities based on their frame of reference (Marsh, 1987). For instance, students in high-ability schools tend to have lower academic self-concepts than students with the same level of ability in low-ability schools (Fang et al., 2018). These phenomena of inaccurate self-evaluations can have unfavourable effects on learning. Overestimation, for example, can be detrimental to learning, because it can elicit underachievement (e.g., Dunlosky & Rawson, 2012). At the same time, the accuracy of

[1] This definition was formulated for the purpose of this work as "accuracy of self-evaluation" tends not to be specifically defined in published studies. However, the definition aligns with the common understanding of the research subject, e.g., Rawson & Dunlosky (2007).

self-evaluation can be developed by learners, and teaching instruction can support this development (e.g., Händel et al., 2020; Naujoks et al., 2022).

Before outlining the structure of this dissertation and describing the main research questions, I would like to elaborate on two important aspects of self-evaluation, as they provide the perspective from which I would like this dissertation to be read and understood. First, I explore the understanding of the "self" in self-evaluation. A mutual concept of self-evaluation seems necessary to relate to the work within this thesis as intended. Second, I address the role of teachers in fostering self-evaluation skills because it highlights the relevance of self-evaluation for themselves and their students at the same time.

A) The Role of the "Self" in Self-Evaluation

In order to explain what is meant by the "self" in self-evaluation, I refer to the feedback model proposed by Hattie and Timperley (2007). Feedback usually applies to a dyadic situation in which one person receives feedback, while the other provides it. If self-evaluation is understood as an internal process – not between two people, but within a single individual – Hattie and Timperley's model of feedback may be adapted to self-evaluation (2007). Self-evaluation may be seen as feedback to oneself. The model suggests that feedback can be given on four levels: the task level, the process level, self-regulation level, and the self level. The first level is called the task level (Hattie & Timperley, 2007). At this level, feedback is provided on the learning task or a learning product (ibid.). The second level refers to the process level and involves feedback given during the learning process, providing the information required to understand or to complete a task (ibid.). The third level describes feedback on self-regulation, such as self-direction or self-discipline (ibid.). The fourth level refers to the self-level (ibid.). The authors emphasise that the self-level was not included for its effectiveness, but rather because of its presence in the classroom and the associated negative consequences (ibid.):

> *"Personal feedback, such as "Good girl" or "Great effort", typically expresses positive (and sometimes negative) evaluations and affect about the student (Brothy, 1981). It usually contains little task-related information and is rarely converted into more engagement, commitment to the learning goals, enhanced self-efficacy or understanding about the task. […] The effects at the self level are too diluted, too often uninformative about performing the task, and too influenced by student's self-concept to be effective. The information has too little value to result in learning gains."*
>
> *Hattie & Timperley, 2007, p. 96*

It is important to note that in this dissertation self-evaluation – viewed as internal feedback – is not understood as feedback on the self level, in line with Hattie and Timberley's criticism of feedback on the self level. Self-evaluation in this dissertation does not relate to personal information or the "goodness" of a person.[2] Self-evaluation does not relate to our own identity, our nature, or ourselves as a person. Instead, self-evaluation is understood as information about a person's behaviour, emotions, or thoughts in relation to goal relevant scholastic criteria. Self-evaluation of learning, therefore, relates to aspects of learning, and nothing more than that. Examples of self-evaluation may include ratings of our understanding, levels of our competence, or learning results – graded or ungraded. States of motivation, frustration, excitement, or effort during learning may also be subjects of our self-evaluation.

Indeed, separating self-evaluation from ourselves can be a challenging thought because it seemingly contradicts how we intuitively perceive self-evaluations. This potentially new perspective describes a mindset that allows us to create a distance between the "self" and the subject of self-evaluation. This detachment is the opposite of over-identification with our learning and may support progress toward our goals, which might otherwise be impeded by an attachment to our "self" as a person. The ideal outcome of this detachment is a neutral, objective attitude towards our learning. To attain this detachment or separation, it seems necessary to define the goals we aim to achieve in learning and the scholastic criteria we apply to judge whether these goals have indeed been met. If we fail to define goals and

[2] For a philosophical perspective on "goodness", see von Wright (1963). Research integrity and good scientific practice may apply these concepts for definitional purposes.

scholastic criteria, self-evaluation may remain “diluted” and “uninformative” as highlighted by Hattie and Timperley (2007, p. 96).

B) Teachers’ Role in Fostering Accuracy of Self-evaluation

Enabling students to become life-long learners is a primary educational goal (UNESCO Institute for Statistics, 2003). Self-regulated learning (SRL) is a key concept in life-long learning, with the accuracy of self-evaluations being a pivotal aspect (see Chapter 1.1 in this dissertation). It needs to be noted that an isolated perspective on students’ self-evaluation without considering teachers’ self-evaluation may remain a constrained perspective. According to an integrative framework in SRL, teachers’ competencies and their instructions influence students’ abilities to self-evaluate their skills (Karlen et al., 2020). However, teachers’ conceptions and practices of SRL are not always aligned with the current scientific understanding of SRL (Dignath & Mevarech, 2021), and their instructions may be adversely affected by their misconceptions. For example, teachers with an autonomy-oriented conceptualisation of SRL seem more likely to use less diagnostic information than those with a motivation-oriented or regulation-oriented conceptualisation of SRL (Dignath & Sprenger, 2020). Moreover, teachers’ beliefs, knowledge, and classroom practices are not always congruent. Even though teachers may have adequate knowledge of metacognition and SRL, discrepancies often arise during the planning and evaluation phases in their instructional practices (Spruce & Bol, 2015). These findings highlight the need to strengthen teachers’ knowledge in SRL and their self-evaluation skills (see also: Kramarski & Kohen, 2017). This dissertation aims to incorporate students’ and prospective teachers’ self-evaluation, emphasizing educators’ responsibility for both themselves and their students.

The motivation to write this dissertation stems from the ideas and thoughts delineated in this introduction. This dissertation is built on a threefold purpose: enhancing the accuracy of self-evaluations, improving self-regulation skills, and fostering a deeper understanding of ourselves as life-long learners – both as educators and students – while detaching our “self”, our identity from the object of evaluation.

The present dissertation is structured as follows: It begins with a theoretical background outlining the role of self-evaluation within self-regulated learning and metacognition – both in general and specifically in biology education –, its formation from a cognitive psychology perspective, and its role in Attention Deficit Hyperactivity Disorder

(ADHS; Chapter 1). The theoretical background and empirical evidence lead to the aim of this dissertation and the main research questions (Chapter 2). The dissertation comprises two published studies and one unpublished manuscript. Each study is briefly summarised, the author’s individual contribution is outlined, and the original manuscripts are included (Chapter 3.1, 3.2, and 3.3). The three studies are discussed in light of the overarching research questions in the general discussion (Chapter 4.1). The methodological approaches are critically discussed (Chapter 4.2). The implications and relevance of the findings are outlined (Chapter 4.3), and the role of non-significant research results is addressed (Chapter 4.4). This dissertation ends with a conclusion (Chapter 5).

1 Self-Evaluation in Education

1.1 Self-Evaluation in Self-Regulated Learning and Metacognition

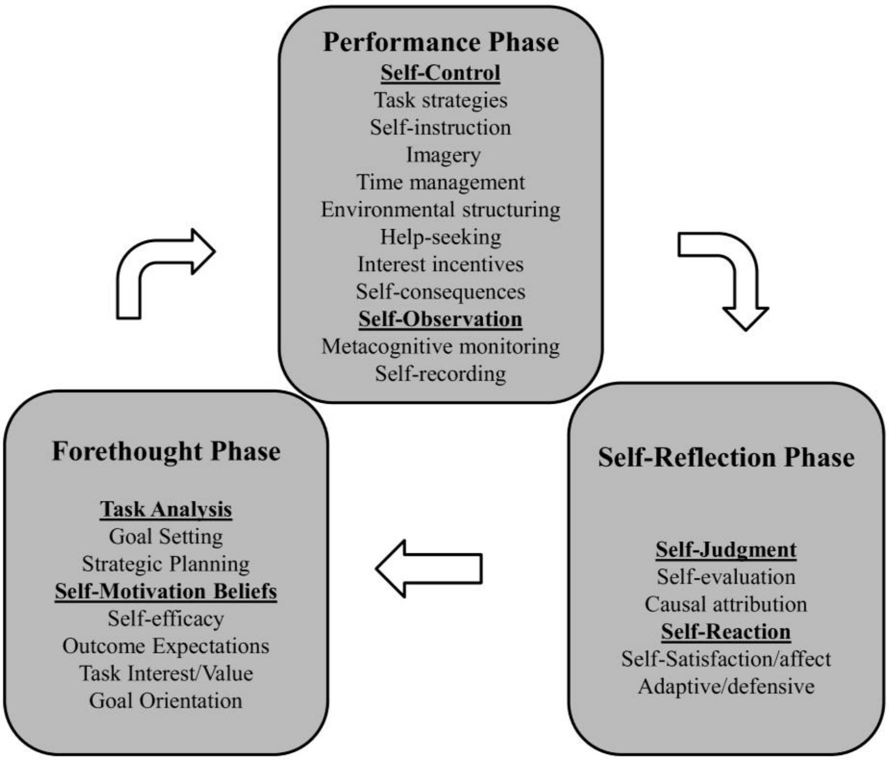
Self-regulated learning (SRL) is one of the key concepts used to support life-long learning (see, for example, Taranto & Buchanan, 2020): an educational objective that has received increased attention following the release of the European Framework of Life-long Learning (Council Resolution on lifelong learning, 2002). This framework demands the provision of "access to life-long learning opportunities for all" and the improvement of "education and training of teachers and trainers involved in lifelong learning" (Council Resolution on lifelong learning, 2002, p. C 162/2). Fostering life-long learning and teaching the necessary skills to support life-long learning are of considerable importance in both educational theory and practice. Understanding the theoretical approaches to life-long learning and the links between self-evaluation and these approaches may contribute to promoting life-long learning in practice.

Various definitions of SRL exist, while most researchers and practitioners focus on SRL as both an individual's ability and as a process. The term SRL is often used to describe an individuals' ability to actively engage in the learning process and guide their own learning (Zimmerman, 1990). SRL may be understood as a multicomponent, iterative, and self-steering process (Boekaerts & Cascallar, 2006; Boekaerts et al., 2005). This process includes cognition, emotions, actions, and environmental factors, which are modulated to serve one's own goals (Boekaerts & Cascallar, 2006; Boekaerts et al., 2005). Numerous models of SRL have been proposed. The six most frequently applied models of SRL have fairly recently been reviewed (Panadero, 2017). These six models are: the cyclical phase model by Zimmerman (e.g., Zimmerman & Moylan, 2009), the dual processing self-regulation model from Boekaerts (e.g., Boekaerts, 2011), Winne and Hadwin's model of self-regulation (e.g., Winne, 2011), Printrich's SRL model (e.g. Pintrich, 2000), the Metacognitive and Affective Model of Self-Regulated learning (MASRL; e.g., Efklides, 2011), and the shared-regulation of learning model in collaborative learning from Hadwin & her colleagues (e.g., Hadwin et al., 2017).

Self-evaluation plays a role in each of these models, for instance, as cognitive judgment in Pintrich's SRL model (2000; as cited in Panadero, 2017) or as an evaluation of goal achievement in the dual processing self-regulation model from Boekaerts (Boekarts and Cascallar, 2006; as cited in Panadero, 2017). Even though self-evaluation is an integral component in each of these models, the cyclical phase model by Zimmerman was selected as the basis for the present dissertation because it offers a comprehensible framework that is easily applicable in practice, was among the first SRL models to be published, and has been widely cited (Panadero, 2017). How self-evaluation is integrated into SRL will be described in more detail in the context of the cyclical phase model by Zimmerman; see Figure 1, adopted from Panadero (2017) citing Zimmerman & Moylan (2009). According to the cyclical phase model, SRL progresses through three phases: the forethought phase, the performance phase, and the self-reflection phase (ibid.). The forethought phase includes the analysis of a given task, including goal setting and planning (ibid.). This phase also incorporates prerequisites for learning activities, including self-efficacy, outcome expectations, and goal orientation (ibid.). The forethought phase transitions into the performance phase (ibid.). Learning activities are executed during the performance phase (ibid.). Strategies of self-control and self-observation, such as self-instruction, time management, and metacognitive monitoring, are applied (ibid.). The performance phase transitions into the self-reflection phase (ibid.). The self-reflection phase involves self-evaluation and causal attribution, summarised as self-judgment and self-satisfaction, as well as affect with adaption and defense, summarised as self-reactions (ibid.).

Figure 1

Cyclical Phase Model of Self-Regulated Learning



Note. This figure was adopted from Panadero (2017, p. 5), citing Zimmerman and Moylan, 2009.

The allocation of these processes is not comprehensive in its entirety, and other constructs and learning strategies in education and related fields may also qualify as strategies of SRL. Nevertheless, self-evaluation has been ascribed considerable relevance in the cyclical process of SRL (e.g., Zimmerman & Moylan, 2009). Self-evaluation is realised in the self-reflection phase and, ideally, results in a transition to the forethought phase (ibid.). The accuracy of self-evaluation may be considered as a prerequisite for this re-entry (ibid.). Based on the model, if self-evaluation is indeed accurate, active learning behaviour may resume, and the cyclical nature of the learning process will be maintained. Inaccurate self-evaluation, on the other hand, such as overestimation, may lead to the termination of active learning behaviour. This may occur if learners feel confident they have achieved a learning goal, even if they have not, in fact, succeeded. It logically follows that learners cease applying active learning strategies. Consequently, the cyclical process of SRL is interrupted.

A relationship between self-evaluation and learning behaviour is not only suggested in SRL theory, but it is also supported by empirical evidence from text reading and item-by-item learning in laboratory studies (Dunlosky & Thiede, 2004; Metcalfe, 2002; Metcalfe & Kornell, 2003; Thiede et al., 2003).[3] Self-evaluation has been shown to be causally linked to learning behaviour, i.e., choosing a number of items to be restudied (Metcalfe & Finn, 2008). Self-evaluation was manipulated, resulting in lower and higher levels of confidence in remembering word pairs in the future, whereas performance remained comparable. Lower levels of confidence lead to choosing more items for restudy than higher levels of confidence, even though item recall performance after learning was equivalent. The conclusion is that self-evaluation directly determines study choices, independently of actual performance (Metcalfe & Finn, 2008). These results suggest that the accuracy of self-evaluation, i.e., the alignment between subjectively perceived and "objectively" parameters of learning, is necessary for adapting appropriate learning behaviour.

It needs to be noted that the cited study is not based on SRL theories. Instead, self-evaluation is examined in the context of metacognition. The theories of metacognition and SRL seem to overlap. Indeed, their conceptualisations are similar (Pintrich et al., 2000). SRL and metacognition were introduced into educational psychology through two distinct lines of research. SRL began to emerge as a widely studied topic in the early 1990s with the publication of the cyclical phase model, one of the first models in this field (Zimmerman, 1989). The term "metacognition" appeared earlier in educational research, introduced by Flavell's definition (Flavell, 1979). Metacognition was defined in the 1970s as "knowledge and cognition about cognitive phenomena" (Flavell, 1979, *p*. 906). Within this publication, metacognitive knowledge and metacognitive experience were distinguished (Flavell, 1979). Metacognitive knowledge refers to knowledge about person, task, and strategy variables in the context of learning. Metacognitive experience is described as "metacognitive knowledge that has entered consciousness" (Flavell, 1979, p. 908). Since the 1970s, the conceptualisation of metacognition has evolved. More recent work distinguishes between declarative

[3] The relationship between self-evaluation and learning behaviour is only briefly described in this dissertation. Nevertheless, two important theoretical approaches need to be mentioned: the discrepancy-reduction model (Dunlosky & Thiede, 2004; Thiede et al., 2003) and the region of proximal learning (Metcalfe & Kornell, 2003; Kornell & Metcalfe, 2006). Both theoretical approaches aim to explain how study time is allocated based on the self-evaluating of learning. They suggest a close relationship between self-evaluation and subsequent learning behaviour and explain how learning can be most effective under varying circumstances. For example, the region of proximal learning framework suggests that there is region of learning "just beyond the grasp of the learner that is most amenable to learning" (Metcalfe & Kornell, 2003, p. 350).

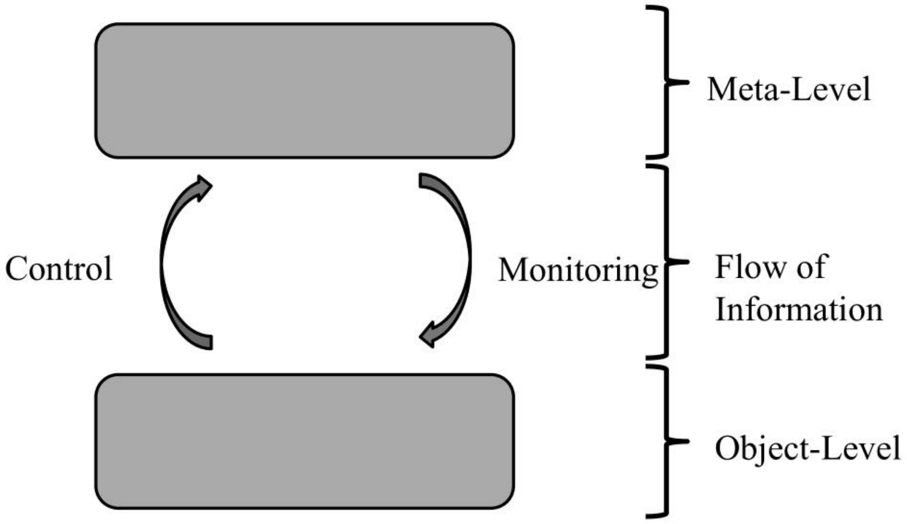
metacognitive knowledge and procedural metacognitive skills (Veenman et al., 2006). Additionally, three principles of metacognition were proposed, describing metacognitive processes in simplified terms (see also Figure 2, Nelson & Narens, 1990):

1. *“Cognitive processes are split into two or more specific interrelated levels,”*
2. *“The meta-level contains a dynamic model (e.g., a mental simulation) of the object level”, and*
3. *“There are two dominance relations, called ‘control’ and ‘monitoring,’ which are defined in terms of the direction of the flow of information between the meta-level and the object-level. “*

Nelson & Narens, 1990, p. 125 - 127

Figure 2

Model of Metacognition



Note. This figure was adopted from Nelson & Narens, 1990, p. 126

Metacognition and SRL exhibit many similarities. For example, SRL includes the use of learning strategies such as imagery in Zimmerman's model, which may be understood as a metacognitive skill according to Flavell's definition. Metacognition involves observing and regulating (Dunlosky & Metcalfe, 2008), which aligns with the phases of the cyclical phase model. Similarly, self-evaluation plays a key role in the reflection phase of the cyclical phase model and in recent research in the field of metacognition (e.g., Metcalfe, 2009), where self-evaluation is most frequently understood as a metacognitive skill. The difficulty of differentiating between SRL and metacognition has already been addressed and remains unresolved (e.g., Pintrich et al., 2000). Metacognition may still be understood to encompass SRL, while SRL may also be understood to encompass metacognition (Pintrich et al., 2000; Veenman et al., 2006). In this dissertation, SRL is considered the broader, overarching concept that incorporates metacognition.[4] Investigating self-evaluation as an aspect of metacognition and SRL may provide a better understanding of life-long learning across various subjects, including biology education.

[4] Motivation, affect, behaviour, and context are aspects of considerable importance in SRL processes (Veenman et al., 2006). These aspects receive less attention in this dissertation due to limitations of scope.

1.2 Metacognition and Self-Evaluation in Biology Education

Already in 2011, the American Association for the Advancement of Science (AAAS) proposed a shift towards active learning in biology education in its report *Vision and Change: A Call to Action* (AAAS, 2011). Following this report, Kimberley Tanner published her widely cited work on metacognition in biology education (Tanner, 2012). She cites the final report, arguing that biology education may greatly benefit from intentional metacognitive instruction. Metacognition as a concept has not specifically been developed in biology education, but as an overarching educational concept (e.g., Flavell, 1979). Nevertheless, it may contribute to a better understanding of the subject biology. Potential benefits of applying metacognition may be differentiated in (A) general benefits and (B) benefits of metacognition that are specific to biology education.

A) Potential Benefits of Metacognition in General

Metacognition is used as an instructional practice to support awareness of one's own learning and to enhance learning performance (for a review in science education, see: Zohar & Barzilai, 2013; for even more comprehensive, general work see: Hacker et al., 2009). While metacognition as a concept is applied, it offers a range of flexible, methodological approaches. First of all, metacognitive instructions may be applied to any field of interest, because planning, execution and evaluation of a task are not per se bound to a specific content. For example, metacognitive strategies such as regulatory checklists or evaluation matrices (Schraw, 1998), are not restricted to any particular topic. Moreover, metacognitive strategies can be applied by learners of any skill level, including both experienced learners and less experienced learners (e.g., Veenman et al., 2006). Importantly, teachers may apply metacognitive strategies not only to improve their content knowledge but also their teaching skills, for instance, through self-questioning. Questions such as "What are my goals for this class session?", "How is the pace of the class going?", or "What evidence do I have that students in my course learned what I think they learned?" are proposed to support metacognition about teaching (Tanner, 2012). It is important to note, that metacognitive instructions may only be effective if the following three principles are realised (Veenman et al., 2006):

1.) Metacognitive strategies need to be embedded in content knowledge.
2.) Learners need to be informed about the usefulness of metacognitive strategies.
3.) Metacognitive strategies need to be continually practiced.

Metacognitive skills may even be transferred to learning tasks that differ from the original task (see far transfer; e.g., Schuster et al., 2020). However, the effectiveness of specific metacognitive strategies remains partially unclear in many areas, including science education (Zohar & Barzilai, 2013). Indeed, metacognitive strategies are not universally effective. For example, a study observed that only approximately half of the university students prompted to use metacognitive skills actually executed their plans (Stanton et al., 2015). Another study with school students in general science classes found that metacognitive strategies were only effective when cognitive training itself was ineffective (Leopold & Leutner, 2015). In a third study, the authors investigated the link between self-evaluation and learning performance. They examined why the accuracy of self-evaluation does not necessarily lead to improved learning, and proposed the contingent-efficacy hypothesis (Dunlosky et al., 2021). The hypothesis posits that the accuracy of self-evaluation does not improve learning when

> (a) *“restudy itself produces only small learning gains for items that were restudied”,*
>
> (b) *“few (or most) of the items have been learned prior to restudy”,* or
>
> (c) *“learners use their accurate judgments inappropriately to make restudy selections*”
>
> *Dunlosky et al., 2021, p. 104*

These examples highlight the importance of identifying both effective and ineffective metacognitive strategies and examining the conditions under which metacognitive strategies may be most beneficial.

B) Potential Benefits of Metacognition Specific to Biology Education

Metacognitive strategies can be beneficial in any educational field. At the same time, biology education is characterised by some specific features with particular usefulness of metacognitive strategies. For example, systems thinking is an inherent aspect of biology education, not only to undergraduate students but across all educational levels. Systems thinking is defined as “a way of thinking to explain, understand, and interpret complex and dynamic (biological) systems” (Evagorou et al., 2009 as cited in Verhoeff et al., 2018, *p.* 2). It describes an understanding of “multiple levels of organization, e.g., molecule, cell, organ, organism, and population, on which phenomena and processes occur and can be explained”

(Verhoeff et al., 2018, p.1). This type of understanding is seen as professional skill (German: “Sachkompetenz”) and required by the scholastic standards set by the Standing Conference of the Ministers of Education and Cultural Affairs in Germany (“Kultusministerkonferenz”; KMK, 2020). Systems thinking seems especially relevant in addressing current global challenges that rely on an understanding of biological and natural science concepts, such as the climate crisis. The complexity of such topics necessitates correspondingly complex learning strategies. Metacognitive strategies can provide the necessary flexibility. By definition, metacognition allows learners to take a bird’s-eye perspective – from the meta-level to the object level – and supports them in their comprehension of biological concepts. Learners may “zoom” into smaller units, e.g., molecules and cells, while maintaining an overview of the superordinate concepts, e.g., organisms and populations. Systems thinking can be assessed and fostered using concept maps, a graphical method that is used in both the practice and research of biology education (Brandstädter et al., 2012).

Metacognitive strategies may also be beneficial for teaching complex methodological skills in biology education. Methods in biology education often require comprehensive understanding of procedures, detailed protocols, and procedural knowledge. Examples of complex methods in biology teaching include chromatography, western plotting, and agarose gel electrophoresis (Reinnard, 2021). Protocol steps are typically precisely defined and numerous. Learners need to be aware of the step-by-step experimental process, and protocols must be carefully followed because even minor mistakes (e.g., choosing the wrong buffer for a polymerase chain reaction or pipetting the incorrect amount of solution) can inadvertently affect the experimental outcome. Metacognitive strategies may support learners in their methodological skill development, e.g., through a step-by-step visualization of the experimental process.

Biological content knowledge is characterised by the use of a variety of technical terms. Examples include terms such as “epigenetics” and “methylation” in genetics, “pelagic zone” in ecology, and “carpel” in botany. Such terms are often applied exclusively in biological contexts and related natural sciences. Remembering these terms and understanding their conceptual meanings may be challenging. This may be particularly true for terms in cellular or molecular biology, as the smallest units cannot be observed with the naked eye. Metacognitive strategies, such as questioning your own understanding of technical terms, can reveal knowledge gaps and provide a basis to address them.

Beyond these theoretical assumptions, a growing body of research has indeed investigated the concept of metacognition in science education (for an overview, see, e.g., Zohar & Barzilai, 2013). Most studies in biology education have focused on university students in higher education (Martin et al., 2000; Palennari, 2016; Sabel et al., 2017; Sebesta & Bray Speth, 2017; Stanton et al., 2019; Stanton et al., 2024; Stanton et al., 2015; Stanton et al., 2021). Several studies have investigated metacognition in scholastic settings (Conner, 2007; McCarthy et al., 2018; Peters & Kitsantas, 2010). Among these studies, self-evaluation is a commonly examined topic. For example, the existence of the Dunning-Kruger effect in biology education has been demonstrated in at least two studies (Osterhage et al., 2019; Ziegler & Montplaisir, 2014). The lowest-performing students displayed the most inaccurate self-evaluations in a university-level biology course (Osterhage et al., 2019). In another study, university students in the upper quartile demonstrated greater accuracy in self-evaluation compared to those in the lower quartile (Ziegler & Montplaisir, 2014). It has also been shown that some biology university students have little experience with self-evaluation (Dye & Stanton, 2017). Specific instructions appear to be necessary to engage with metacognitive strategies productively (Sabel et al., 2017). Senior students, as well as introductory, students use information about their performance to evaluate their plans (Stanton et al., 2019). Some students use emotions (Stanton et al., 2019). These results illustrate that inaccuracies in self-evaluation are also evident in biology education. Encouragingly, the accuracy of self-evaluation can indeed be improved. For instance, specific teaching instructions, including self-evaluation strategies (Osterhage et al., 2019), or repeated practice of self-evaluation (Ziegler & Montplaisir, 2014) have been shown to enhance the accuracy of self-evaluation. A combination of metacognitive training, psychoeducation, feedback, and the use of specific judgements have also been shown to reduce the effects of overestimation (Händel et al., 2020). To understand how educational practices can further foster the accuracy of self-evaluation, it is essential to consider how self-evaluation is formed.

1.3 Formation of Self-Evaluation

The formation of self-evaluation is subject to a range of cognitive processes. In an attempt to explain how self-evaluations are derived, two classes of theories have been proposed: direct-access and inferential theories (Nelson et al., 1984; Schwartz, 2024). Direct-access theories suggest that self-evaluative judgements are formed directly based on memory strength (Schwartz, 2024). According to this approach, self-evaluation and actual performance should be closely aligned, as they rely on the same source of information (Schwartz, 2024). However, this has been questioned by empirical evidence showing that self-evaluation and actual performance are not necessarily aligned, and self-evaluation may be inaccurate (e.g., Koriat & Bjork, 2005). Inferential theories challenge the direct-access approach (Schwartz, 2024). Central to inferential theories is the idea that self-evaluation of one's own learning cannot be directly derived from the memory representation of studied material, which is usually declarative in nature. Instead, self-evaluation relies on various cognitive processes (e.g., Thiede et al., 2005). The cue-utilization framework, an inferential theory, outlines what these different cognitive processes may involve (Koriat, 1997).

The cue-utilization framework explains how self-evaluative judgments during and after learning may be formed prior to retrieval (Koriat, 1997). According to the cue-utilization framework, cues are used to estimate one's own learning (see Table 1 for empirically tested cues and Table 1 in Study II of this dissertation for a theoretical categorisation).

Table 1

Overview of Cues Used for Self-Evaluation

Cues used for self-evaluation	**Empirical study**
Learners' belief about own general memory efficacy	Hertzog et al. (1990)
Characteristics of study situation	Begg et al. (1989), Zechmeister & Shaughnessy (1980)
Type of expected memory test	Mazzoni & Cornoldi (1993)
Previous task-specific experience	Hertzog et al. (1990), King et al. (1980), Mazzoni & Cornoldi, 1993; Schneider (1986)
Perceived relative difficulty of study items	Arbuckle & Cuddy (1969)

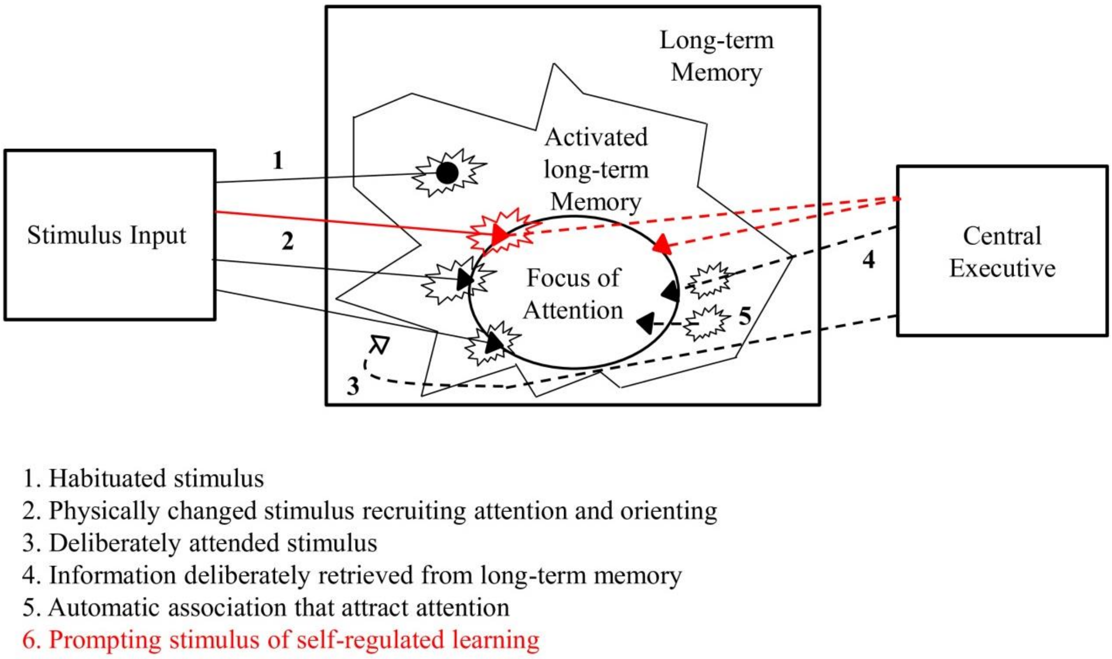
Note. This table was adopted from Koriat (1997).

Three types of cues may be used: intrinsic, extrinsic, and mnemonic cues (Koriat, 1997). Intrinsic cues include item characteristics that serve as predictors for the difficulty of an item (ibid.). These cues are inherent attributes of the study material (ibid.). Extrinsic cues refer to characteristics of the learning conditions and encoding processes during learning (ibid.). Such cues may include the number of study repetitions or a learner's level of processing (ibid.). Mnemonic cues are described as internal cues (ibid.). An internal cue may include the ease with which information comes to mind or familiarity with a cue (ibid.). These three types of cues may impact self-evaluation directly or directly by impacting other types of cues (ibid.). The cue-utilization framework provides a framework that describes self-evaluation as flexible and adaptable to the external and internal environment (ibid.). It describes the formation of self-evaluation as a highly complex process, also because the number of cues than may be used is not constrained (ibid.). Any information encountered during the learning process may serve as a cue for the formation of self-evaluation. At the same time, the human cognitive system has limited capacity, and not all incoming information is processed at the same level (see, for example, model of working memory from Baddeley, 1992). The concept of working memory enhances our understanding of the cognitive processes involved in self-evaluation. In a general definition, working memory refers to "the system or systems that are assumed to be necessary in order to keep things in mind while performing complex tasks such as reasoning, comprehension, and learning" (Baddeley, 1992, p. R136). Multiple models of working memory have been proposed (see, for example, Miyake & Shah, 1999). This dissertation relates to Cowan's embedded-processes model of working memory (Cowan, 1988, 1999) because it focuses on functions such as attention and stimulus processing rather than features such as the episodic buffer and the phonological loop in Baddeley and Hitch's working memory model (Cowan et al., 2020; Repovš & Baddeley, 2006).[5] Cowan emphasises constraints in working memory capacity and defines it as follows: "The ensemble of components of the mind that hold a limited amount of information temporarily in a heightened state of availability for use in ongoing information processing." (Cowan, 2017, as cited in Cowan et al., 2020). The embedded-processes model consists of multiple components and describes the functions among them. It describes how information may be processed so that it can be used to perform a task (Cowan, 1988, 1999). The model is illustrated in Figure 3.

[5] For an overview of working memory models and a comparison with the embedded-processes model, see Cowan, N., Morey, C. C., & Naveh-Benjamin, M. (2020). An embedded-processes approach to working memory. *Working Memory: The state of the science, 44.*

Figure 3

The Embedded-Processes Model of Working Memory


Long-term Memory
Activated long-term Memory
Stimulus Input
Focus of Attention
Central Executive
1
2
3
4
5
1. Habituated stimulus
2. Physically changed stimulus recruiting attention and orienting
3. Deliberately attended stimulus
4. Information deliberately retrieved from long-term memory
5. Automatic association that attract attention
6. Prompting stimulus of self-regulated learning

Note. This figure was adapted from Cowan et al., 2020, p. 48.

The embedded-processes model of working memory comprises a sensory store, long-term memory, activated long-term memory, the focus of attention, and central executive processes (Cowan, 1988, 1999). The sensory store is the first component that holds information. It retains information for only a few hundred milliseconds (ibid.). The sensory store can activate features of the long-term memory system (ibid.). Only a small proportion of these activated elements can be in the focus of attention and be used to perform a task (ibid). The focus of attention is also time-limited, but information may be held for longer compared to the sensory store (ibid.). Incoming and recently (and deliberately) attended information can remain in an active state in long-term memory for a time that is not clearly defined (ibid.). Information in the active state is characterised by a heightened state of availability (ibid.). This information is more accessible and may more easily attain the status of focuses attention (ibid.). Information may become inactive if it is not well-consolidated, whereas well-consolidated information may stay active for longer (ibid). Processes that control attention are classified as central executive processes (ibid.). Central executive processes are associated with goals and the aim to achieve these goals (ibid.). These processes are considered to be deliberate (ibid.).

According to the embedded-processes model of working memory, stimuli from the external and internal environment are processed (ibid.). These stimuli may follow several processing paths (ibid.). Habituated stimuli (1) are external stimuli that may become part of activated long-term memory, but do not attain the status of focused attention (ibid.). External stimuli may attain the status of the focus of attention and may be used for task engagement (2), e.g., words written on a blackboard in school (ibid.). Stimuli may also be deliberately attended to and purposely attain the status of focused attention (3), e.g., arguments in a peer discussion that a learner may decide to pay attention to (ibid.). Stimuli may also be retrieved from long-term memory and attain the status of focused attention (4), e.g., an already well-understood biological concept that may be integrated into a concept map (ibid.). There are also automatic associations in long-term memory that can attract attention (5) (ibid). Attention is essential to the processes of the working memory (ibid.). Attention has long been investigated in psychological research, but there is still a lack of a common definition. William James described attention as follows:

> *"Everyone knows what attention is. It is taking possession by the mind in clear and vivid form, of one out of what seem several simultaneously possible objects or trains of thought. Focalization, concentration, of consciousness are of its essence. It implies withdrawal from some things in order to deal effectively with others, and it is a condition which has real opposite to the confused, dazed, scatterbrained state which in French is called distraction, and Zerstreutheit in German."*
>
> *William James, 1890*

The embedded-processes model suggests that information can only be utilised for the execution of a task if it obtains the status of focused attention (Cowan, 1999). Attention may be deliberately directed by the learner (see central executive processes in the embedded-processes model of working memory) or guided by external sources. Instructional practices in education are processes through which learners' attention is deliberately and purposefully guided. This deliberate guidance is particularly important for the formation of self-evaluation because self-evaluation relies on cues about learning (see cue-utilization framework; Koriat, 1997). However, not all information arising from these cues is predictive of learning performance. Cues that are less predictive of actual performance are associated with decreased accuracy of self-evaluation (Prinz-Weiß et al., 2023; Serra & Dunlosky, 2010; Thiede et al., 2010), and learners may not use the most predictive cues intuitively, which are associated with more accurate self-evaluation. Guiding learners' attention towards more predictive cues during the learning process seems beneficial, e.g., focussing on one's own ability to explain a text rather than on the quality of a text itself (Thiede et al., 2010). Which cues are more predictive than others remains a subject of ongoing research and is partially addressed in this dissertation (see Study II).

Deliberate guidance can be used flexibly: prior to learning, during learning, and after learning, in accordance with the phases of self-regulated learning: the forethought phase, the performance phase, and the self-reflection phase (Zimmerman, 1990). Ideally, deliberate guidance of attention succeeds not only in activating focus of attention, but also in stimulating central executive processes. By stimulating central executive processes, learners may be able to self-regulate their own cognitive processes (see Figure 3, red markings). A prompt that may help stimulate central executive processes could be: "Please rate your comprehension of the text on a scale from 0 to 100%. Describe how you know how well you understood the text. What do you think could help you improve your judgement?"

Although educators and teachers can deliberately guide their students' learning based on the concepts of cue-utilization and the embedded-processes model of working memory (see: Cowan, 1999; Koriat, 1997), there may still be specific circumstances under which deliberate guidance seems particularly difficult, for instance, when learners' lack fundamental skills to direct their own attention. Attention deficit hyperactivity disorder (ADHD) is the most commonly diagnosed developmental disorder in childhood and adolescence and is characterised by distinct impairments in directing one's own attentional processes (Willcutt, 2012). Self-evaluation and ADHD are addressed in the following paragraph.

1.4 Self-Evaluation and ADHD

Attention Deficit Hyperactivity Disorder (ADHD) is defined as a neurodevelopmental disorder of self-regulation (Schlottke et al., 2019). It is diagnosed based on three core symptoms (Diagnostic and Statistical Manual of Mental Disorders; DSM-5; American Psychiatric Association, 2013). These core symptoms are inattentiveness, impulsivity, and hyperactivity (ibid.). Three subtypes of ADHD are distinguished: the combined subtype, characterised by symptoms of inattention and hyperactivity-impulsivity; the predominantly inattentive subtype, with symptoms of inattention but not hyperactivity-impulsivity; and the predominantly hyperactive-impulsive subtype, with symptoms of hyperactivity-impulsivity but not inattention (ibid.). The predominantly inattentive subtype is also referred to as Attention Deficit Disorder (ADD, e.g., Marshall et al., 1997). To be diagnosed with ADHD, the symptoms must be present for at least six months, and the individual's quality of life must be impaired (Faraone et al., 2021). Among children and adolescence, the prevalence rate ranges from 6.1 to 9.4 % (Salari et al., 2023). No difference in prevalence rates is observed between countries and regions when controlling for the use of different diagnostic tools (Willcutt, 2012). Boys are diagnosed with ADHD twice as often as girls (ibid.). Symptoms of ADHD may persist into adulthood (Faraone et al., 2006). Fifteen percent of children diagnosed with ADHD continue to fully meet the diagnostic criteria at the age of 25 (ibid.). Sixty-five percent partially meet the diagnostic criteria (ibid.). The prevalence rate among young adults is approximately 5% (Willcutt, 2012).

The developmental disorder ADHD is recognised as a disorder of self-regulation (Schlottke et al., 2019). Because this dysfunction of self-regulation is also present in learning settings, ADHD is of particular relevance in the context of metacognition and self-regulated learning (SRL). ADHD is associated with cognitive and emotional impairments, including difficulties in regulating attention and motivation (e.g., Sonuga-Barke, 2003). Symptoms of ADHD strongly and negatively impact academic achievement (Arnold et al., 2020), and academic underachievement is frequently observed. Learners with ADHD appear less likely to reach their scholastic potential (Barry et al., 2002; Kent et al., 2011). In detail, children with ADHD work less persistently, tend not to complete their homework, apply less effort, and used more superficial learning strategies (Hoza et al., 2001; Langberg et al., 2016; O'Neill & Douglas, 1991).

While many aspects of self-regulated learning may be affected in ADHD, the impact on self-evaluation may be one of the key aspects. The positive illusory bias (PBI) is a phenomenon observed in children with ADHD describing overly positive self-evaluations

(Hoza et al., 2002). Although methodological issues in PBI research arise, for instance, absolute self-perceptions or a lack of objective measurements of competence (or a comprehensive and critical review see: Owens et al., 2007), findings indicating overly positive evaluations in children with ADHD appear robust. For example, individual studies have shown that boys with ADHD overestimated their scholastic competence, their social acceptance, and behavioural conduct compared to boys without ADHD based on teacher ratings (Chan & Martinussen, 2016; Hoza et al., 2002). Furthermore, children with the combined subtype with symptoms of inattention and hyperactivity-impulsivity overestimated their math achievement compared to children without ADHD (Owens & Hoza, 2003). Interestingly, these children also overestimated their reading and mathematical skills compared to children with the inattentive subtype (Owens & Hoza, 2003). More severe hyperactivity-impulsivity symptoms are associated with the positive illusory bias but not with more severe inattention (Owens & Hoza, 2003). Moreover, other types of inaccuracies in self-evaluation occur in children with ADHD. For example, overly positive self-evaluations were associated with externalizing problems (Volz-Sidiropoulou et al., 2016), and children with ADHD seem to attribute success to luck rather than their own skills (Hoza et al., 2001).

Nevertheless, children with ADHD do not generally seem to make inaccurate evaluations, and improvements in self-evaluations in children with ADHD seem an achievable goal. For example, while the evaluation of their own skills is prone to errors, their peers' skills are accurately evaluated by children with ADHD (Evangelista et al., 2008). Additionally, the type of question may prompt children with ADHD to more accurate responses of self-evaluation, e.g., specific ratings lead to more accurate skill estimation than global ratings (Prevatt et al., 2012). Deliberate strategies and approaches designed to enhance the accuracy of self-evaluation may be particularly beneficial for children with ADHD.

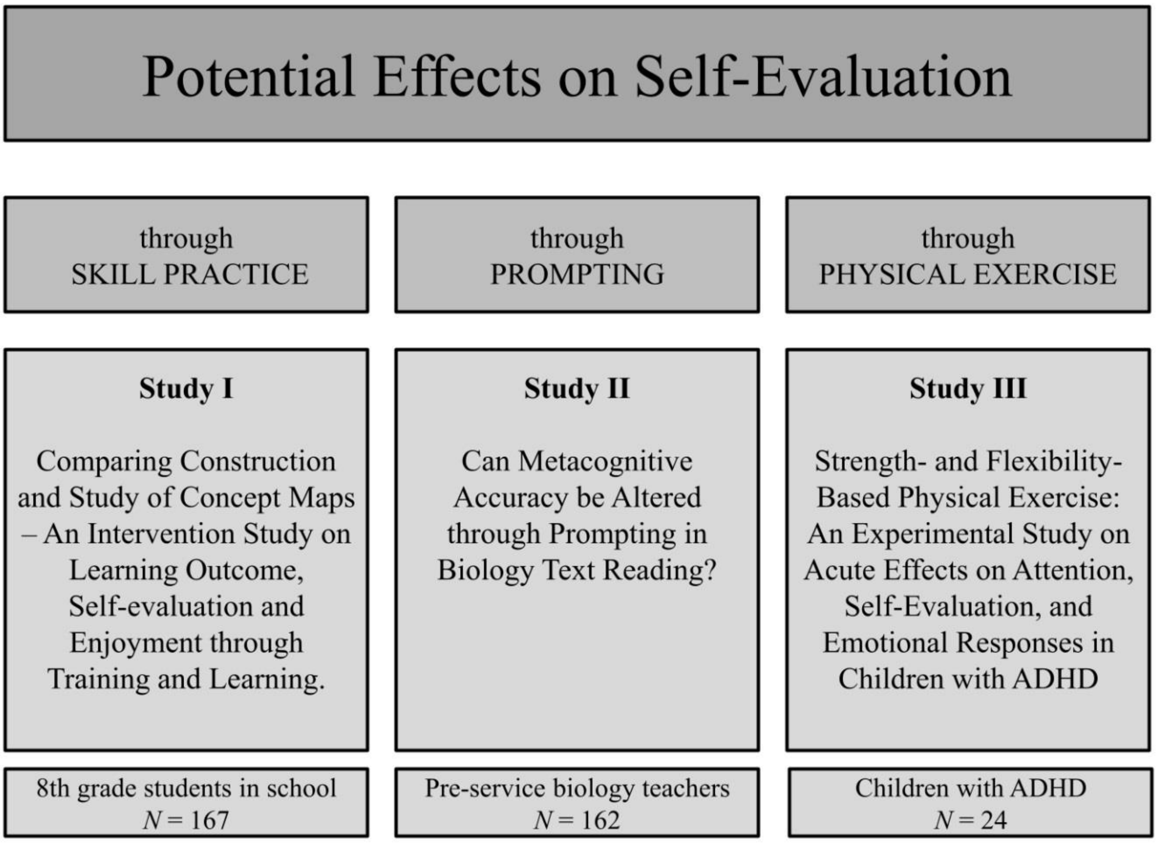
Chapter 1 delineates the role of self-evaluation in self-regulated learning and metacognition, its formation from a cognitive psychology perspective, and its role in ADHD. The dissertation aim and the study overview are presented in the next Chapter.

2 Dissertation Aim and Study Overview

The main aim of this dissertation is to examine potential effects of different approaches on the accuracy of self-evaluation. Three studies were planned, conducted, and analysed within this dissertation (see Figure 4). The first study examined whether the accuracy of self-evaluation can be improved through the practice of the skill that is to be evaluated. For this purpose, 167 eighth-grade students took part in an intervention study. The second study examined whether the accuracy of self-evaluation can be improved through prompting during text reading. For this purpose, 162 pre-service biology teachers took part in an online study. The third study examined whether the accuracy of self-evaluation can be improved through acute physiological changes induced by physical exercise. For this purpose, 24 children with Attention Deficit Hyperactivity Disorder (ADHD) participated in an experimental laboratory within-subjects study.

Figure 4

Overview of the Studies included in this Dissertation



Potential Effects on Self-Evaluation

through SKILL PRACTICE

Study I

Comparing Construction and Study of Concept Maps – An Intervention Study on Learning Outcome, Self-evaluation and Enjoyment through Training and Learning.

8th grade students in school
N = 167

through PROMPTING

Study II

Can Metacognitive Accuracy be Altered through Prompting in Biology Text Reading?

Pre-service biology teachers
N = 162

through PHYSICAL EXERCISE

Study III

Strength- and Flexibility-Based Physical Exercise: An Experimental Study on Acute Effects on Attention, Self-Evaluation, and Emotional Responses in Children with ADHD

Children with ADHD
N = 24

3 Empirical Studies

Three studies are included into this dissertation. These studies will be briefly summarised. Research questions and methodological approaches will also be outlined. The author's contribution will be declared.

3.1 Study I: Promoting Self-Evaluation through Skill Practice

Study I was published in 2022 in the *Frontiers in Education*. Sina Lenski and Stefanie Elsner share first authorship. The study can be found under the following citation:

> Lenski S., Elsner S. & Großschedl J. (2022). Comparing Construction and Study of Concept Maps – An Intervention Study on Learning Outcome, Self-Evaluation and Enjoyment Through Training and Learning. *Frontiers in Education.* 7:892312.doi: https://doi.org/10.3389/feduc.2022.892312

3.1.1 Study I: Summary, Research Questions and Methodological Approach

Study I aims to answer the question of whether the accuracy of self-evaluation can be improved through the development of the skill being evaluated. Previous research has shown that self-evaluation can be improved through enhancing the skill that is being evaluated (Kruger & Dunning, 1999). However, this result has been shown in adults and in non-academic contexts. The present study aims to extend these findings to skills related to learning strategies in biology education in schools. It is important to note that the present study, included in this dissertation, was not specifically designed to answer only this particular research question. Moreover, the study was conducted to examine the overall effectiveness of different concept map training approaches. The question of how the training impacts self-evaluation is one of several research questions. The study's main research questions and results will be outlined, with a particular emphasis on self-evaluation.

A recently published meta-analysis examined the effects of using concept maps on the learning process (Schroeder et al., 2018). This meta-analysis distinguished between the effects on learning when concept maps were either constructed or studied without construction. This meta-analysis provided important insights into the most effective ways to learn with concept maps. However, methodological differences between single studies included may limit implications, as studies typically investigate either the construction or the study of concept maps separately. To my knowledge, no study has directly compared the effects of constructing versus studying concept maps. The present study included in this dissertation aims to bridge this gap. Effects of concept map training, including construction and concept map study, were investigated. This study addresses the following research questions[6]:

> **RQ 1**: Do training lessons in concept map construction and study affect *learning outcomes*?
>
> **RQ 2**: Do training lessons in concept map construction and study affect the *quality of concept maps*?
>
> **RQ 3: Do training lessons in concept map construction and study affect *self-evaluation*?**
>
> **RQ 4:** Do training lessons in concept map construction and study influence students' enjoyment of learning with concept maps?
>
> **RQ 5:** Are the effects of concept map training transferable to learning with concept maps of the other type? That is, does learning by studying concept maps benefit from prior training in concept map construction, and vice versa?

To address these research questions, an intervention study with 167 eighth-grade students was conducted. A quasi-experimental, 3x2-factor design with a cross-over of training type and learning type was applied. To determine whether trainings in the construction and study of concept maps affected self-evaluation, Spearman correlations were calculated. For this, self-evaluated concept mapping skills and the "objective" assessment of concept map quality were

[6] The present study also examined the effects on cognitive load as an important dependent variable. However, due to scope limitations, the results related to cognitive load will receive less attention in this dissertation.

correlated. The limitations of this method in determining the accuracy of self-evaluation will be discussed in Chapter 4.2.

3.1.2 Study I: Own Contribution

The author of this dissertation carried out essential aspects of data preparation, data analysis, data visualization, and manuscript writing. The author also prepared the data for open-access publication through the Open Science Framework (DOI: https://osf.io/mw356/).

The study was designed by Sina Lenski, co-author with shared first authorship, and Jörg Großschedl, listed as third author. The study was conducted by Sina Lenski and five student assistants as part of their theses within the project. Jörg Großschedl revised manuscript drafts and provided supervision. The authors' contributions can also be found in the original manuscript.

3.1.3 Study I: Published Manuscript

Comparing Construction and Study of Concept Maps – An Intervention Study on Learning Outcome, Self-Evaluation and Enjoyment Through Training and Learning

Lenski, S., Elsner, S., & Großschedl, J.

Abstract

Concept maps are graphical tools for organizing and representing knowledge. They are recommended for biology learning to support conceptual thinking. In this study, we compare concept map construction (CM-c, i.e., creating concept maps) and concept map study (CM-s, i.e., observing concept maps). Existing theories and indirect empirical evidence suggest distinct effects of both formats on cognitive, metacognitive and emotional aspects of learning. We developed a CM-c training, a CM-s training, and a brief introduction to concept maps (control training) for junior high school students. We investigated effects on *learning performance*, *concept map quality*, *cognitive load* (cognitive effects), accuracy of *self-evaluation* (metacognitive effects) and *enjoyment* (emotional effects) of these trainings in a subsequent learning phase (CM-c learning vs. CM-s learning) in a quasi-experimental two-factorial study with 3 x 2 groups (*N* = 167), involving the factors training type and learning type. Results reveal that CM-c training increased *learning performance* and *concept map quality*. Effects of CM-c training on *learning performance* transferred onto learning with CM-s. *Self-evaluation* was slightly more accurate after CM-c training than CM-s training. Students reported moderate and highly varying *enjoyment* during CM-c and CM-s learning. The superiority of CM-c over CM-s in *learning performance* and *concept map quality* probably lies in its characteristic of being an active learning strategy. We recommend practitioners to favor CM-c training over CM-s training, and foster students' active engagement and *enjoyment*.

Keywords: Biology education, concept maps, metacognition, learning performance, training

Introduction

Natural sciences deal with the description, explanation and prediction of natural phenomena. Inherent to understanding the natural sciences is conceptual thinking. Conceptual thinking involves organization of new knowledge and the integration of it into already existing knowledge. Modern biology lessons aim to provide opportunities for students to develop skills in conceptual thinking, and educate students to apply these skills to become solution-focused problem solvers. While conceptual thinking can be challenging for students (OECD, 2016; Ekinci and Şen, 2020), it can be encouraged through many different learning strategies. Working with concept maps provides such a learning strategy (e.g., Tseng, 2020). Concept maps (CMs) are network-like diagrams for organizing and representing knowledge. They summarise and visualise the most important concepts of a topic and the relationships between these concepts. Concepts are linked with labeled arrows whereas the direction of the arrowheads specify the reading direction. Concept map construction (CM-c) is the process of creating a concept map (mostly) based on textual material by self-organizing concepts and arrows. Concept map study (CM-s), on the other hand, is the process of viewing a previously designed (expert-)concept map without additional textual material.

Concept maps have been intensively examined and further developed since their introduction in the 1970s by Joseph Novak. Many recommendations were given for their use (see e.g., Schroeder et al., 2018 for a recent overview). Heterogeneous results regarding the learning effectiveness of concept maps are often explained by the notion that the learners had different expertise in the use of concept maps. Up to now, it is controversially discussed whether concept map training is necessary in order to use concept maps successfully and how this training should be structured. While previous studies primarily focused on cognitive aspects of learning with concept maps (e.g., learning performance and concept map quality), metacognitive and emotional aspects have scarcely been addressed. However, learning processes are generally accompanied by metacognitive and emotional activities (e.g., self-evaluation and enjoyment) whilst directly or indirectly influencing learning outcome.

This study presents and examines two concept map trainings, focusing on concept map construction on the one hand and concept map study on the other. The aim of this study was to (1) develop a training structure based on theoretical foundation and empirical evidence, (2) examine aspects of cognitive, metacognitive, and emotional effects of familiarity with concept maps on the learning process, and (3) investigate to what extent expertise with one learning format (e.g., concept map study) is conducive to the use of the other format (here:

concept map construction). We specifically aim at deriving implications for practitioners and future research from our study.

Theoretical Framework

Learning Effectiveness of the Construction and Study of Concept Maps

CM-c and CM-s are regularly used in classrooms and empirical comparison of their effects on learning seems valuable. Learning with concept maps can yield improved learning outcome (Visible Learning Meta^{X} Research Base ®, 2021). This is especially prevalent when CM-c and CM-s are compared with other learning strategies. Learners who constructed concept maps outperformed learners who took notes (Reader and Hammond, 1994), created summaries, discussed with fellow students (Chularut and DeBacker, 2004), marked texts (Amer, 1994), and read texts or attended a lecture (Nesbit and Adesope, 2006; Woldeamanuel et al., 2020; Hwang et al., 2021). Learners who studied (animated) concept maps outperformed others who studied texts (Rewey et al., 1989; Patterson et al., 1992; O´Donnell et al., 2002; Nesbit and Adesope, 2011), lists (Lambiotte et al., 1993), or outlines (Salata, 1999). Meta-analyses report mixed findings when comparing CM-c and CM-s based on effect sizes. Horton et al. (1993) observed greater benefits for CM-s than for CM-c. In contrast, Adesope and Nesbit (2013) and Schroeder et al. (2018) observed greater benefits for CM-c than CM-s. The more recent meta-analysis including more studies and larger sample sizes, provide evidence for superiority of CM-c over CM-s in *learning performance*. We are not aware of empirical studies that directly compared the effects of CM-c and CM-s on learning outcome. Comparing CM-c and CM-s will offer insight into the robustness of theory-driven cognitive mechanisms of learning with concept maps. Findings might also provide guidance for practitioners to make decisions about learning strategy use.

Cognitive Effectiveness of the Construction and Study of Concept Maps

Based on Ausubel’s theory on learning (Ausubel et al., 1978), it is argued that concept maps promote meaningful learning (Novak and Cañas, 2008; Schroeder et al., 2018). Meaningful learning is taking place when new knowledge is created or assimilated into existing interconnected knowledge structures through cognitive elaboration (Novak and Cañas, 2008). Meaningful learning involves well-organized, relevant knowledge structure and emotional commitment to integrate new knowledge with existing knowledge (Novak and Cañas, 2008). Potential cognitive effects of learning with concept maps are proposed (Nesbit and Adesope, 2006; Schroeder et al., 2018). They include: (1) Dual coding through visual and verbal information in concept maps supports effective retrieval, (2) Cognitive load is reduced and overloading of the memory system is prevented, (3) Centralization of the key concept allows for better semantic integration, (4) Semantic structure is marked more clearly compared to text formats, (5) Simple syntax allows for easy access to learners with yet poor reading and writing abilities, (6) Greater elaborative thinking is promoted through decision making processes, and (7) Greater elaborative thinking is promoted through higher degree of concision and summarization.

With respect to these proposed cognitive effects, a distinction must be made between different concept map formats. CM-c and CM-s differ particularly in their degree of elaborative thinking and cognitive load (mechanisms 2, 6, and 7). CM-c is presumed to promote learners' active engagement with the interconnections of the content (Hardy and Stadelhofer, 2006; Freeman et al., 2014); it is more cognitively demanding, supports deeper engagement, and fosters a higher level of elaborative thinking than CM-s (Schroeder et al., 2018). Taken together, enhanced *learning performance* through CM-c than CM-s can be assumed. The impact on other relevant learning variables is likely to differ between CM-c and CM-s, too.

Construction and Study of Concept Maps –Training, Cognitive Load, and Transfer

Despite a small number of studies concluding that a short introduction to concept maps is sufficient or that learning with concept maps does not need to be practiced at all (Ruiz-Primo, 2004; Ifenthaler, 2011; Karpicke and Blunt, 2011), research predominantly recommends concept map practice. Most scholars in the field support the notion that the learning effectiveness of concept maps depends on the degree of familiarity with this learning method (Holley and Dansereau, 1984; Renkl and Nückles, 2006; Correia et al., 2008; Mintzes et al., 2011; Aguiar and Correia, 2017; Großschedl and Tröbst, 2018). Trainings (i.e., extended periods of practice) increase familiarity and hence support learning effectiveness. It was shown that CM-c trainings improve the ability to construct concept maps (den Elzen-Rump and Leutner, 2007; Jin & Wong, 2010; Sumfleth et al., 2010; Leopold and Leutner, 2015; Becker et al., 2021). In line with this, it was observed that expertise in the use of knowledge maps (Chmielewski and Dansereau, 1998) and concept maps (Chang et al., 2002) improves knowledge structuring and information encoding when summarizing texts. CM-s training increased level of expertise measured through eye movement (Lenski and Großschedl, im Druck). For untrained students, on the other side, CM-c yielded negative effects on *learning performance* (Neuroth, 2007).

These negative effects are probably due to excessive cognitive load. Learners' working memory may get overloaded when processing two types of information simultaneously: strategy-related information about concept mapping and learning-related information about learning contents. Learners might experience a so-called *map shock* when studying concept maps. This is characterized by "bewilderment of not knowing where to start or how to penetrate the topography of the map" (Blankenship and Dansereau, 2000; p. 294).

Theoretically, memory resources can be occupied by three types of *cognitive load: intrinsic*, *germane,* and *extraneous load* (Sweller, 2010). *Intrinsic load* arises from the difficulty and complexity of the task. It depends on the number of interacting elements (element interactivity) and learners' prior knowledge. *Intrinsic load* can be manipulated by activating the learners' prior knowledge or simplifying the learning content (Klepsch and Seufert, 2020).

Intrinsic load cannot be altered directly by the design of learning material. On the other side, *extraneous load* is caused by suboptimal design of learning material (e.g., plain, text-based learning materials; e.g., Poppenk et al., 2010; Orru and Longo, 2018). A reduction in *extraneous load* could free resources to be available for acquiring and automating schemes in long-term memory (*germane load*). *Germane load* refers to the learning-related load and comprises resources that are available for acquiring and automating schemes in long-term memory.

Increasing the familiarity with concept maps through training could result in a reduction of *intrinsic* and *extraneous load; and prevent a map shock*. Greater familiarity with the task could reduce the amount of new strategy-related information, simplify the learning process, and reduce the perceived difficulty (intrinsic load, Young et al., 2014). As a consequence, more cognitive resources for content-related processes (germane load) will be available (Mayer and Moreno, 2003).

We presume *intrinsic* (H1.3a) and *extraneous cognitive load* (H1.3b) to be reduced and *germane load* (H1.3c) to be increased through both, CM-c training and CM-s training. We expect this effect to be evident compared to a control training. Furthermore, we assume that learners who are trained in the use of CM-c or CM-s, show improved skills in constructing concept maps (*concept map quality*) (H1.2) and increased *learning performance* compared to untrained learners (H1.1a).

We additionally aim at understanding whether skills acquired through training in one specific format of working with concept maps impact working with another format. Although both learning formats are somewhat similar, it needs to be assumed that different skills are needed for each type of learning e.g., CM-c learning requires learners to (re-)structure, CM-s learning requires learners to recognize information and compare new knowledge with already existing knowledge. We address the question whether CM-c training is conducive to CM-s and vice versa. If such a transfer effect exists, we might see similar results in *learning performance* when learning with CM-c and CM-s after CM-c training. We assume that CM-c training has higher transfer potential on CM-s learning than CM-s training has on CM-c learning, because concept mapping skills are probably transferred from the (more) active type of use to the (more) passive type of use (H1.1b). Taken together, an advantage of CM-c training on cognitive measurements is expected.

Metacognition in Concept Map Trainings: Accuracy of Self-Evaluation

The accuracy of *self-evaluation* refers to the congruency of "objective" and subjective performance evaluation. *Self-evaluation* is conceptually placed within the frameworks of metacognition and self-regulation (see Flavell, 1979; Panadero, 2017). Both frameworks refer to abilities that include planning, monitoring, and evaluating one's own learning processes (Schraw, 1998; Panadero., 2017). Metacognition emphasizes the observer's perspective and is described as 'thinking about thinking' (Flavell, 1979). One's own thoughts become objects of thoughts themselves. Accuracy of *self-evaluation* is placed within the evaluation aspect of self-regulation and metacognition.

Accuracy of *self-evaluation* is pivotal when practicing a new learning strategy, because it might determine appropriate adjustment of learning efforts towards a learning goal. Following Zimmerman's idea of a circular learning process (Zimmerman, 2000) accurate *self-evaluation* leads to adapted planning behavior. This means, high congruency of *self-evaluation* results in more appropriate planning behavior by students and goal attainment of the learning goal becomes more likely. However, accurate *self-evaluation* is not always naturally existent. Empirical studies suggest that some students overestimate, and others underestimate their abilities in various skills (Kruger and Dunning, 1999). The Kruger-Dunning effect was shown to be less evident after improving these skills (Kruger and Dunning, 1999). We assume that the Kruger-Dunning effect probably occurs in working with concept maps as well, and can be overcome by CM training. Through CM trainings, students acquire necessary declarative and procedural skills. Hence, student's ability to accurately *self-evaluate* their own skills is likely to improve. While we assume that both trainings (CM-c and CM-s) improve student's *self-evaluation,* we expect higher accuracy following a CM-c training (H2). We expect this because of a higher degree of procedural concept map experience in CM-c training.

Emotion in Concept Trainings: Enjoyment

According to Ausubel et al. (1978), emotional commitment is an inherent part of meaningful learning. Emotional commitment to a learning task is reflected in the construct of *enjoyment*. *Enjoyment* can be defined as an activity related affective state (Pekrun et al., 2006). It is experienced when the activity or the learning material is positively valued and perceived as controllable by the learner (Pekrun et al., 2006). Experiencing *enjoyment* increases task engagement and supports persistent use of a learning strategy beyond training or a formal research study. A few studies report insights into the perception of *enjoyment* during concept map tasks. Romero et al. (2017) observed that students largely enjoy working with concept maps. Percentages of 77.8 and 88.2% of two groups of 13 to14 year old students stated to “like working on the subject through concept mapping experience”. A study with university students indicates that *enjoyment* differs between learning formats (Blunt and Karpicke, 2014). Students gave higher reports of *enjoymen*t for constructing concept maps after reading a text compared to summarizing the same text in a paragraph (while the text is still present). In this study moderate *enjoyment* was reported (29 to 51 on a scale from 0 = “not at all” to 100 = “totally”).

CM-trainings have the potential to increase *enjoyment*. Negative affective states which accompany (potential) excessive cognitive demands might be reduced as a consequence of familiarity with concept maps. Learners will be more likely to perceive the task as controllable. We assume that CM-c and CM-s trainings increase familiarity with concept maps, reduce cognitive demands and therefore increase *enjoyment* with working with concept maps. Potential differences between the learning formats (CM-c learning, CM-s learning) are of equal interest in this study.

Overview of the Study

We investigate the effects of concept map trainings (CM-c training, CM-s training, control training) and concept map learning type (CM-c learning, CM-s learning) on cognitive (*learning performance, concept map quality*, *cognitive load*), metacognitive (accuracy of *self-evaluation*) and emotional aspects (*enjoyment*) through a direct comparison.

Based on the theoretical foundation, the following hypotheses arise:

H1.1: We assume that learners who are trained in the use of CM-c or CM-s show increased learning performance compared to untrained learners **(a)**. Furthermore, we assume that CM-c training has higher transfer potential on CM-s learning than CM-s training has on CM-c learning, because concept mapping skills are probably transferred from the (more) active type of use to the (more) passive type of use **(b)**.

H1.2: We hypothesize that learners who are trained in the use of CM-c or CM-s, show improved skills in constructing concept maps (concept map quality).

H1.3: We presume intrinsic **(a)** and extraneous cognitive load **(b)** to be reduced and germane load **(c)** to be increased through both, CM-c training and CM-s training compared to a control training.

H3: We assume that CM-c and CM-s trainings increase familiarity with concept maps, reduce cognitive demands and therefore increase enjoyment with working with concept maps.

Materials and Methods

This study was conducted at non-academic track schools during regular school days and term. One instructor conducted the study in all classes and was assisted by one of three assistants. All assistants received the same instructions and performed the same tasks. Bothe, the instructor and the assistants supported students in case instructions or clarification are needed. We followed the respective local school law agreements (North Rhine-Westphalian Ministry of Education Science and Research, 2005) and the ethical principles and guidelines for the protection of human subjects of research (Department of Health, Education, and Welfare, 2014).

Design and Procedure

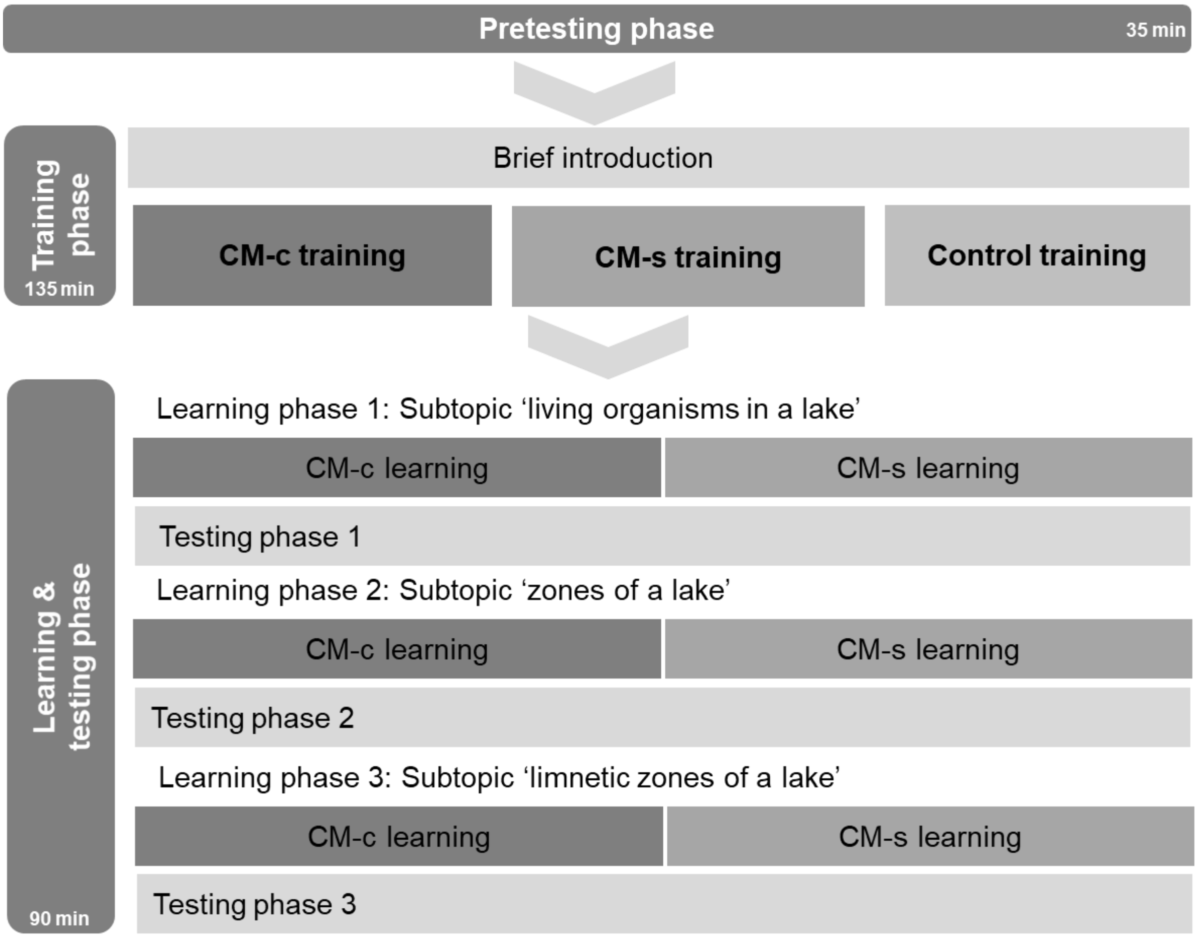
Schools were contacted via e-mail, flyer or personally. Classes were invited to take part in the quasi-experimental intervention study. We received the greatest response from non-academic track schools. The study covered a period of about 3 weeks and was carried out in regular biology or natural science lessons (see Figure 1). The study involved three main phases: firstly, a pretesting phase; secondly, a training phase, and thirdly, a combined learning and testing phase. The entire study comprised six lessons of 45 min each with visiting times of two lessons each week. Pretesting phase, in which demographic data were gathered, took place in the first school lesson. It was identical for all participants. Subsequently, entire classes were randomly assigned to one of the trainings by drawing lots. Entire classes underwent either a CM-c training, a CM-s training or a control training. Training phase lasted for three lessons. After the training phase, students were randomly assigned to either one of two types of learning. Within one class, half of the students studied through CM-c learning and the other half studied through CM-s learning. Students studied with individual workbooks. In this learning and testing phase, students' ability to develop knowledge through CM-learning was measured. A second set of workbooks was used to assess the effects of training and learning. In these textbooks, students provided answers to test questions and variables of interest. Learning and testing phase lasted for two lessons. The stepwise randomization (first step: class level, second step: student level) resulted in a two-factorial design with 3 x 2 groups. Of 58 students that took part in the CM-c training, 31 students studied through CM-c learning and 27 students studied through CM-s learning in the learning and testing phase. Of 59 students that took part in the CM-s training, 29 students studied through CM-c learning and 30 students studied through CM-s learning in the learning and testing phase. Of 50 students that took part in the control training, 20 students studied through CM-c learning and 30 students studied through CM-s learning in the learning and testing phase. Supplementary Material 1 shows resulting groups.

Participants

A total of 201 eighth-graders from nine classes (between 12 and 35 students per class) at non-academic track schools in North Rhine-Westphalia, Germany participated in this study. The 8th grade was chosen because, according to the curriculum, method training can be integrated well here. Supplementary Material 1 gives an overview of participant allocation, exclusion criteria and the variables analyzed. We excluded thirty-four students from data analyses because crucial parts of the study were missed. Eighteen students were excluded because they took part in less than two out of three training sessions. Sixteen students were excluded because they were late for class and missed parts of the learning and testing phase. The remaining *N* = 167 participants were on average *M* = 14.05, *SD* = 0.82 years old. Of all participants, 47.3% were female and 44.9% were male (7.8% did not provide an answer). A percentage of 52.1% were German native speakers and 25.7% stated another language than German as their first language (22.2% did not provide an answer). Reading fluency was lower (80.42 ± 13.62) than in norm samples (100 ± 15) as assessed by Salzburger Lesescreening (Auer et al., 2005). The average biology grade was 2.68 (grading scale from 1 = “very good” to 6 = “insufficient”). Students were informed that this study will not affect their academic reports. In one class, only a small number of students gave evaluable answers to the questions regarding *cognitive load*, *self-evaluation* and *enjoyment* leading to reduced sample sizes for these variables (see Supplementary Material 1). We note that the instruction was disregarded by the students.

Figure 1

Study Design



Pretesting Phase

During pretesting phase, we gathered students' demographic information including age and gender, reading fluency and prior knowledge about ecosystems to account for individual differences potentially influencing *learning performance*. Reading fluency was assessed through the Salzburger Lesescreening 5-8 with reported reliability of r_{tt} = 0.89 (SLS 5-8; Auer et al., 2005). This test measures reading speed and reading comprehension by means of a list of simple sentences. Students are asked to read these sentences as quickly as possible and determine their truthfulness. The test can be assessed in class and takes about 10 minutes to execute. Prior knowledge about ecosystems in general and the ecosystem lake was evaluated in a written test including single and multiple-choice questions (see Supplementary

Material2). The questionnaire consisted of six self-developed questions and two modified questions obtained from Keusch and Telaak (2017). Additionally, three questions were obtained from the third International Mathematics and Science Study TIMSS (Harmon et al., 1997; Baumert et al., 1998), as the items were validated for grade eigth and cover the topic ecosystems (see Supplementary Material 2, items taken from the TIMSS study are marked accordingly). An item on general knowledge about ecosystems includes, for example, the task of filling in an incomplete food chain (Supplementary Material 2, p. 3, item 4). An item focusing on the lake ecosystem covers, for example, the limnetic zone of a lake (Supplementary Material 2, p. 5, item 6). Test scores were transformed into a percentage value with 100% indicating solely correct answers. We report a Cronbach's α of 0.36.

Training phase

All trainings were based on cognitive theories as recommended by Collins et al. (1988), Klauer (1988), and Renkl (2010). The *theory of adaptive control of thought* (ACT; Anderson, 1983) recommends to teaching declarative knowledge (e.g., facts, ideas, and rules) followed by procedural knowledge (knowledge of how an activity is performed) to acquire competence in a certain process. Based on this, all trainings began with a 25-minute introduction to concept mapping. This introduction included declarative knowledge about concept maps, the general idea of concept maps and the use of this new learning method. In CM-c and CM-s trainings, procedural knowledge about CM-c and CM-s was conveyed. The *cognitive apprenticeship theory* (CAT; Collins et al., 1988) is a constructivist approach to instruction. Cognitive and metacognitive processes which take place during the execution of complex tasks are made visible. This is done by an instructor who verbalizes these processes while the task is performed and provides support and feedback for the learners when performing the task on their own.

Based on this, students underwent four phases (*modelling*, *scaffolding*, *fading*, and *coaching*). The modeling phase was administered for declarative introduction (instructor constructs a sample concept map on the blackboard) whereas the remaining three phases were only carried out in the CM-c and the CM-s trainings but not for the control training. Students in the control training did not receive any further instruction or in-depth information on concept maps beyond the 25-min introduction to concept maps. Instead, students took part in a non-academic social training (team building activity) which did not include a learning activity (see Supplementary Material 3 for detailed description of the trainings and their theoretical foundation). In Lenski and Großschedl (2021), the complete teaching concept for the construction training in German including all necessary materials is available.

Learning and testing phase

In the learning and testing phase, we examined students' ability to develop knowledge through CM-c learning and CM-s learning. Students studied the topic "ecosystem lake" in three subtopics ("living organism in a lake", "zones of a lake", "limnetic zones of a lake") through either CM-c learning or CM-s learning. The three subtopics were studied consecutively with a learning period of 20 min each with individual workbooks. During CM-c learning, students constructed concept maps based on learning texts. Stickers with concepts were provided to promote and simplify the construction of concept maps (for a similar approach see Gehl, 2013). During CM-s learning, students were asked to study expert designed concept maps. These concept maps had been designed based on the same textual material as used in CM-c learning. Validity was secured through three independent raters with content equivalence of o Fleiss' $\kappa = 0.96$ for concept map 1 ("living organisms in a lake"), of Fleiss'$\kappa = 1$ for concept map 2 ("zones of a lake"), and of Fleiss'$\kappa = 0.82$ for concept map 3 ("limnetic zones of a lake").

After students studied each subtopic, we measured *learning performance, concept map quality* (only for CM-c learning, not CM-s learning)*, cognitive load, self-evaluation,* and *enjoyment.* This resulted in three measurements for all variables providing more valid data than one measurement.

Instruments

Learning performance

We assessed learning performance on the topic ecosystem lake by a paper-based questionnaire with open-ended and single choice questions. The questionnaire can be obtained from Supplementary Material 4. This questionnaire comprised five self-developed questions, two questions from the TIMSS study (Harmon et al., 1997) and 16 modified questions based on Keusch and Telaak (2017). Test scores were transformed into a percentage value with 100% indicating solely correct answers. We report internal consistency of Cronbach's $\alpha = 0.75$.

Concept map quality

We assessed *concept map quality* through a scoring system as suggested by Clausen and Christian (2012). It allows evaluation of concept map structure and content. Students in CM-c learning condition constructed three concept maps on three subtopics of the “ecosystem lake”. Numbers between one and five were assigned for each proposition accounting for the type of relation, labels and connecting structures; 0 = two linked concepts without substantial relation, 1 = two linked concepts, arrow without label but with substantial relation, 2 = two linked concepts with labeled arrow and descriptive relation, 3 = two linked concepts with hierarchical relation, 4 = cause-effect relation without labeled arrow, 5 = cause-effect relation with labeled arrow. Numbers were added to a sum-score. Two rating teams evaluated ten percent of all maps while one rating team rated the entire material. We report an interrater reliability of Cohen’s κ = 0.75 for concept map 1 (“living organisms in a lake”), of Cohen’s κ = 0.94 for concept map 2 (“zones of a lake”), and of Cohen’s κ = 0.94 for concept map 3 (“limnetic zones of a lake”). One overall mean value of all three concept map-sum-scores was calculated for each student.

Cognitive load

We assessed *cognitive load* via the seven-item version of a self-reporting questionnaire designed by Klepsch et al. (2017). We measured *extraneous* (*ECL*), *intrinsic* (*ICL*) and *germane load* (*GCL*). Questionnaire statements were modified only by the replacement of “the task” with “the concept map” (e.g., “When looking at concept maps, many things needed to be kept in mind simultaneously.”). Students rated statements on a 7-point Likert scale ranging from “*I fully disagree*” to “*I fully agree.*” Mean values for the subscales over all three times of assessments were computed. We report the following internal consistencies: *extraneous load* (*ECL*, Cronbach’s α = 0.68 - 0.78), *intrinsic load* (*ICL*, Cronbach’s α = 0.55 - 0.75), *germane* load (*GCL*, Cronbach’s α = 0.75 - 0.78).

Self-evaluation

Self-evaluation on students’ concept map skills was measured with five statements; “I read the text thoroughly,” “I used all the concept stickers,” “I paid attention to the direction of the arrows.”, “I labelled all the arrows.” and “I understood connections between concepts.” Students rated their agreement on a three-stepped emoticon-based scale (joyful, indifferent, sad smiley) according to den Elzen-Rump and Leutner (2007). We report internal consistencies for *self-evaluation* for each subtopic (concept map 1: Cronbach's $\alpha = 0.68$, concept map 2: Cronbach's $\alpha = 0.77$, concept map 3: Cronbach's $\alpha = 0.76$).

Enjoyment

Enjoyment was measured with a single question in reference to Blunt and Karpicke (2014). *Enjoyment* was measured three times after each of the three learning periods ("living organism in a lake," "zones of a lake," "limnetic zones of a lake"). We asked students to answer the question "How much did you enjoy this task?" on a written scale from 0 to 100% in increments of 10%.

Preliminary tests and statistical analyses

Preliminary tests were carried out at an α-level of 0.10 to determine potentially existing differences between training groups before students' participation in the intervention. Choosing an α-level of 0.10 allows to indirectly minimize the β-error in statistical analyses in which the null hypothesis is "favored". The null hypothesis is "favored" in preliminary tests because we assume no differences between training groups at baseline. One-way analyses of variance (ANOVAs) and a chi-square test were carried out. Results indicated that there were no differences between training groups in reading fluency, $F(2,130) = 2.04, p = 0.135$, prior knowledge about ecosystems, $F(2,152) = 0.76, p = 0.471$ or gender proportions, $\chi^2(2) = 1.34, p = 0.513$ but in age, $F(2,152) = 2.98, p = 0.054$ (for descriptive data see supplementary Material 5). As we perceive reading fluency and prior knowledge as greater predictors of *learning performance* than age, we did not regard the age difference between training groups as substantial. For most variables, analyses on standard distribution and outliers (> 3× interquartile range) did not yield unusual data distribution. Alternative tests were used in the case of a violation of assumptions (see section "Results" for specific tests applied).

Throughout the results section we use the terms "TRAINING" and "LEARNING" for the two independent variables. "TRAINING" relates to the type of training, which students took part in: CM-c training, CM-s training, control training. "LEARNING" relates to the type of learning phase, which students underwent subsequently to training. Students studied either through CM-c or CM-s. All main hypotheses were tested at an α-level of 0.05. We applied two-way analysis of variances to investigate differences in *learning performance* and *enjoyment* through CM training and learning (H1.1a.b; H3). We ran one-way analyses of variances to determine differences in *concept map quality* between training groups (H1.2).

We used two-way multivariate analyses of variances to investigate differences in *cognitive load* (resp. *extraneous*, *intrinsic*, *germane cognitive load)* through CM training and learning (H1.3a – c). Bonferroni corrections were applied as *post hoc* analyses for statistically significant results following analyses of variances. We ran Spearman correlations for ordinal data with *self-evaluation* and *concept map quality* to determine accuracy of *self-evaluation* (H2). Correlations allow us to determine congruency of two variables with each other. If not provided by IBM SPSS Statistics (version 24.0), effect sizes were calculated according to Lenhard and Lenhard (2016). Because of missing data in the control group and potential distorting statistical results, we interpret statistical results for *cognitiv load*, *self-evaluation* and *enjoyment* in both training groups but not in the control group.

Results

Learning performance

To investigate whether training type (CM-c training, CM-s training, control training) and type of learning (CM-c learning, CM-s learning) influenced *learning performance*, we ran a two-way analysis of variance on *learning performance*. Table 1 and Figure 2 show means and standard deviations of *learning performance*.

Table 1

Means and standard deviations for learning performance, concept map quality, cognitive load, self-evaluation and enjoyment separate for training type and learning and testing phase

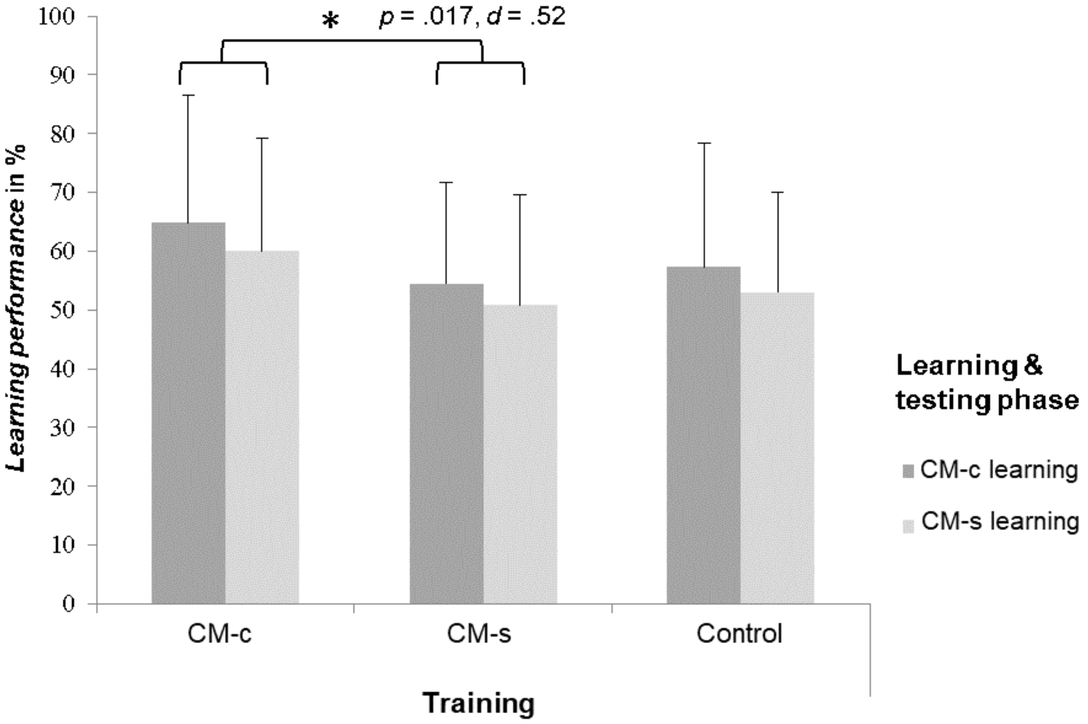
Training Type	**CM-c training**						**CM-s training**						**Control training**					
Learning and testing phase	**CM-c learning**			**CM-s learning**			**CM-c learning**			**CM-s learning**			**CM-c learning**			**CM-s learning**		
	M	***SD***	***n***	***M***	***SD***	***n***	***M***	***SD***	***n***	***M***	***SD***	***n***	***M***	***SD***	***n***	***M***	***SD***	***n***
Learning performance in %	64.78	21.77	31	60.5	19.07	27	54.53	17.08	29	50.83	18.74	30	57.35	21.09	20	52.99	17.12	30
Concept map quality	29.88	15.58	31	n.a.	n.a.	n.a.	20.03	13.90	29	n.a.	n.a.	n.a.	15.77	14.09	20	n.a.	n.a.	n.a.
ECL	3.55	1.35	31	3.36	1.37	27	3.92	1.53	28	3.54	1.26	29	3.50	0.97	7	3.13	1.08	29
GCL	4.24	1.53	31	3.68	1.23	27	4.17	1.50	28	4.31	1.42	29	5.40	0.97	7	3.67	1.48	29
ICL	3.79	1.29	31	4.11	1.17	27	3.82	1.27	28	4.45	1.57	29	5.26	1.05	8	4.18	1.11	29
Self-evaluation	2.54	0.49	30	2.45	0.52	27	2.44	0.42	28	2.56	0.47	29	2.65	0.26	7	2.50	0.43	29
Enjoyment	45.16	35.42	31	26.85	24.95	27	37.70	25.79	29	37.70	25.79	29	67.38	30.15	7	36.67	24.88	29

Note. CM-c, concept map construction; CM-s, concept map study a cognitive load was measured on a seven-point Likert scale ranging from (1) = low cognitive load to (7) = high cognitive load, self-evaluation was measured on a three-stepped pictorial scale, enjoyment was measured on a scale from 0 to 100%

We observed that *learning performance* was higher after CM-c training (62.58 ± 20.52%) compared to CM-s training (52.65 ± 17.89%); $F_{TRAINING}$ (2, 161) = 4.03, $p = 0.020$, $\eta^2_p = 0.05$ with *post hoc* analyses (Bonferroni) resulting in $p = 0.017$, $d = .52$ (partially support for H1.1a). We observed no differences between the control training (54.74 ± 18.73%) and both CM trainings ($p_{\text{CM-c training}} = 0.105$; $p_{\text{CM-s training}} = 1.00$). We did not find that the type of learning impacted *learning performance* (CM-c learning: 59.21 ± 20.28%, CM-s learning: 54.44 ± 18.50%); $F_{LEARNING}$ (1,161) = 2.03, $p = 0.157$. We did not observe an interaction of training type with type of learning; $F_{TRAINING\ X\ LEARNING}$ (2, 161) = 0.01, $p = 0.989$ (support for H1.1b).

Figure 2

Means and Standard deviation of Learning Performance



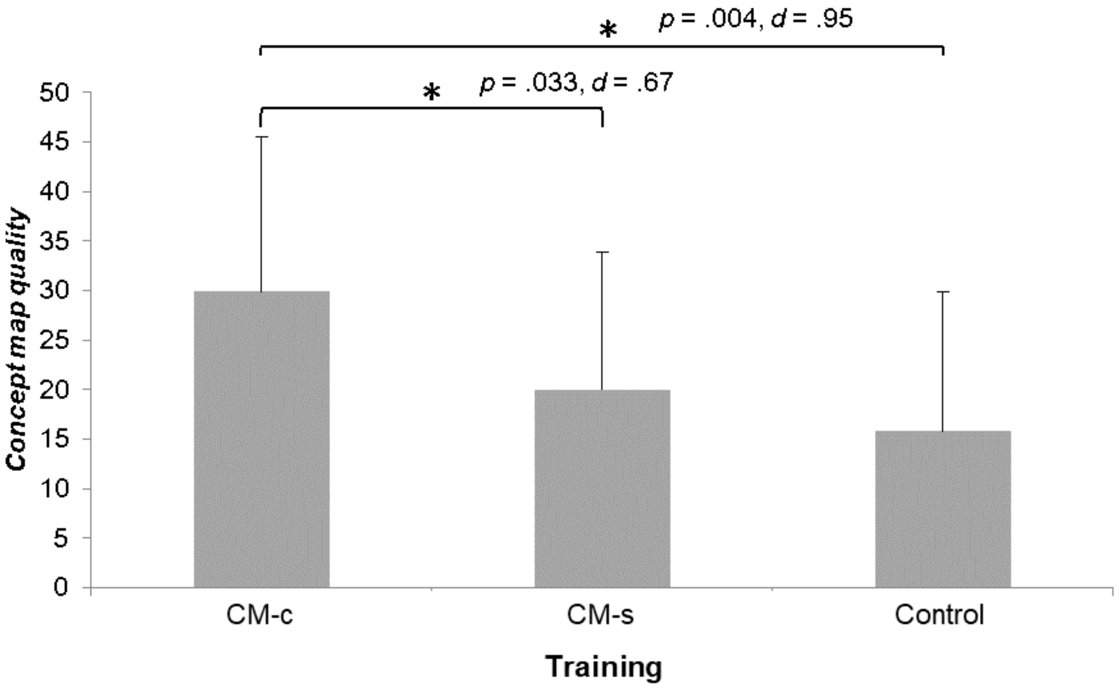
Note. CM-c = concept map construction, CM-s = concept map study

Concept map quality

To examine differences in *concept map quality* between training groups (CM-c training, CM-s training, control training) during CM learning, we ran a one-way analysis of variance. Table 1 and Figure 3 show means and standard deviations for *concept map quality*. Results showed that concept map quality was higher following CM-c training (29.88 ± 15.58) compared to CM-s training (20.03 ± 13.90), $F(2,77) = 6.47$, $p = 0.003$, $\eta^2_p = 0.14$ with *post hoc* analyses (Bonferroni) of $p = 0.033$, $d = 0.67$. *Concept map quality* was also higher following CM-c training compared to the control training (15.77 ± 14.09; $p = 0.004$, $d = 0.95$) (partially support of H1.2). There was no difference between CM-s training and the control training ($p = 0.956$).

Figure 3

Means and Standard deviation of Concept Map Quality



Note. CM-c = concept map construction, CM-s = concept map study

Cognitive load

To investigate whether training type (CM-c training, CM-s training) and type of learning (CM-c learning, CM-s learning) influenced *cognitive load*, we ran a two-way multivariate analysis of variance on *cognitive load* including *extraneous (ECL), intrinsic (ICL)* and *germane load (GCL)).* Table 1 shows means and standard deviations. Results of the multivariate analysis revealed no difference in cognitive load between training groups F_{TRAINING} (3, 109) = 0.45, $p = 0.715$, Wilks' $\Lambda = 0.99$, $\eta^2_{\text{p}} = 0.12$, but a difference between type of learning phase F_{LEARNING} (3, 109) = 5.25, $p = 0.002$, Wilks' $\Lambda = 0.87$, $\eta^2_{\text{p}} = 0.13$. This effect did not reach statistical significance after *post hoc* testing [F_{ICL} (1, 111) = 3.63, $p = 0.059$, $\eta^2_p = .032$; F_{GCL} (1, 111) = 0.61, $p = 0.437$, $\eta^2_{\text{p}} = 0.005$; F_{ECL} (1, 111) = 1.20, $p = 0.277$, $\eta^2_p = 0.011$].

No interaction of training type with type of learning phase was evident $F_{\text{TRAINING x LEARNING}}$ (3, 109) = 1.55, $p = 0.205$. Taken together, training type (CM-c training, CM-s training) and type of learning did not differ in their impact on students' cognitive load (lack of support of H13a-c).

Self-evaluation

We investigated whether CM trainings influenced accuracy of students' *self-evaluation*. In our study, accuracy of *self-evaluation* is reflected in the congruency of students' *self-evaluation* (evaluation of concept map skills) and objective assessment (*concept map quality*). As a measurement of congruency, we ran Spearman correlations for ordinal data with *self-evaluation* and *concept map quality* for each training group. High correlations indicate high accuracy of *self-evaluation*. Correlations reveal highest accuracy after CM-c training ($\text{r}_s = 0.66$, $p < 0.001$, $n = 30$), followed by CM-s training ($\text{r}_s = 0.52$, $p = 0.004$, $n = 28$) and the control training ($\text{r}_s = 0.60$, $p < 0.159$, $n = 7$; partially support for H2). Table 1 shows means and standard deviations for *self-evaluation* and *concept map quality*. We observed that only a small number of participants in the control training provided answers to *self-evaluation* questions. Only a comparison between correlations after CM-c training and CM-s training is legitimate.

Enjoyment

To investigate whether training type (CM-c training, CM-s training) and type of learning (CM-c learning, CM-s learning) influenced emotional commitment to learning with CMs, we ran a two-way analysis of variance on *enjoyment. Enjoyment* was analyzed with Box-Cox transformed data because of a violation of homogeneity of error variances. Table 1 shows untransformed means and standard deviations for *enjoyment*. We observed moderate *enjoyment* and high variability across students (38.35 ± 30.09%) with a range of 0 to 100% in *enjoyment*. Students reported average enjoyment following the CM-c (36.64 ± 32.09%) and CM-s training (37.40 ± 29.33%) with high variability during learning phase. Training type did not influence *enjoyment;* $F_{TRAINING}$ (1, 111) = 0.40, $p = 0.530$ (lack of support for H3). We observed no effect of type of learning; $F_{LEARNING}$ (1, 111) = $2.12 \cdot 10^4$, $p = 0.988$. Training type and type of learning did not interact; $\mathrm{F}_{\mathrm{TRAINING\ X\ LEARNING}}$, $F(1, 111) = 3.26$, $p = 0.074$. It needs to be noted that analyses revealed a violation of the assumption of homogeneity of error variances. Box-Cox transformation reduced heterogeneity but did not entirely stabilize data as assessed by Levene's test, $p = 0.036$. The unusually dispersed data might have obscured potential effects. Results need to be observed and interpreted with caution.

Discussion

Learning performance and concept map quality

As expected, results show higher *learning performance* for students who took part in CM-c training instead of CM-s training (partially support for H1.1b). As we observed that CM-c training improved *concept map quality* (partially support for H1.2), it is likely that the increased *learning performance* is a result of improved concept mapping skills.

In line with other findings (Hilbert and Renkl, 2008; Jin and Wong, 2010; Sumfleth et al., 2010), we assume that CM-s training and the control training are not sufficient to enable students to construct concept maps. A specific training in the construction of concept maps is needed to improve students' ability to construct *concept map* as suggested by other authors (e.g., den Elzen-Rump and Leutner, 2007; Sumfleth et al., 2010; Großschedl and Tröbst, 2018). Students were able to apply these skills and to engage more deeply with the learning content. This finding supports the assumption that CM-c promotes elaborative thinking. Elaborative thinking probably takes place to a greater extent in CM-c than in CM-s. We ascribe this superiority of CM-c training in *learning performance* to its active nature. Active learning tasks are generally associated with increased *learning performance* (McCagg and Dansereau, 1991; Chang et al., 2002; Freeman et al., 2014).

However, contrary to our hypothesis we did not observe a difference in *learning performance* between CM trainings and the control training. We assume that students who took part in the control training probably did not acquire the necessary skills to effectively apply CM-c or CM-s during learning. Instead of applying concept mapping skills, students probably used other learning strategies that appeared to be beneficial for them in the past (e. g. repeated reading) (see Wild, 2001 for more information on individual learning strategy use). This is supported by the observation of lower concept quality after the control training. Increase in *learning performance* following the control training cannot be explained by an increase in concept mapping skills.

In conclusion, in contrast to CM-s training, CM-c training enabled students to apply concept mapping skills to a degree that allowed them to learn effectively with concept maps. Students improved their ability to construct concept maps and they were able to use this learning strategy to acquire similar knowledge as the use of other naïve strategies would. To be able to use concept maps as a more effective way of learning, we suggest practice of more than three lessons. The maximum potential of concept maps as a learning strategy might only be exploited by a prolonged training.

Transfer effect

We addressed the questions whether CM-c training impacts CM-s learning and vice versa. Our results show CM-c training increased *learning performance* irrespective of whether students constructed or studied concept maps in a subsequent learning task (support for H1.1b). Here, the absence of a statistically significant interaction effect suggests the existence of a transfer effect. An evident interaction effect (i.e., higher *learning performance* after CM-c training for those students who constructed concept maps during learning and testing phase but not for those students who studied concept maps) would have suggested that skills learned through CM-c training are only applied in CM-c learning but not in CM-s learning. We did not observe such an interaction effect and conclude that skills learned through CM-c training are also applied in CM-s learning. The CM-c training most likely altered student’s overall information processing strategies, enabling them to implicitly interact with a different CM learning format. This is in line with previous studies suggesting that the familiarity with particular formats can positively influence *learning performance* in similar formats (e.g., Royer and Cable, 1976; Royer, 1979). Our results could be explained by the nature of the tasks (passive vs. active learning task). The familiarity in an active learning task (here CM-c) has higher transfer potential compared to the passive learning task. We conclude that CM-c training benefits *learning performance* regardless of which learning format (CM-c or CM-s) is applied after training.

Cognitive load

We expected *intrinsic* (H1.3a) and *extraneous cognitive load* (H1.3b) to be reduced and *germane load* (H1.3c) to be increased through both, CM-c training and CM-s training compared to the control training. Statistical results showed that CM-c training and CM-s training did not differ in their impact on *cognitive load*. We observed no difference between types of learning.

That cognitive load seemed uninfluenced by training in our study, reflects methodological limitation instead of providing an answer to our research question. We surmise that the used instrument did not differentiate between sources of *ECL* and *ICL* as mentioned by Klepsch and Seufert (2020), which was published after the conduction of this study. For settings where *ICL* and *EGL* may be intertwined, Klepsch and Seufert (2020) recommend using complex instruments to uncover the underlying processes. We also suspect methodological issues with measuring *GCL* and agree with the authors of the instrument that the "wording of the current items was ambiguous, so learners understood them differently" (Klepsch et al., 2017, p. 9). Therefore, our findings should be treated with caution. Further research is needed to find measurements that reliably assess *cognitive load* during learning activities. We emphasize that simple and clear language that is comprehensible also for low-achieving students should be used.

Self-evaluation

We assumed that CM trainings increase accuracy of *self-evaluation* while we expected that CM-c training has higher influence than CM-s training. Our data only allow a comparison of CM-c and CM-s because of a low number of participants in the control group. Based on effect sizes, results show that accuracy of *self-evaluation* is improved through CM-c training to a greater extent than CM-s training (partially support for H2). We assume that this outcome is due to higher amount of procedural knowledge was shown by the statistical significant difference in *concept map quality* after CM-c and CM-s training (H1.2). Beyond this, we would like to address the question whether accurate *self-evaluation* is a premise or a consequence of successful skill acquisition. The answer to this question has relevant implications for practitioners. If accurate *self-evaluation* is a premise, teachers should include teaching methods that support *self-evaluation* such as providing opportunities for students to reflect on their current level of task skills. If accurate *self-evaluation* is a consequence of successful skill acquisition, teachers should focus on students' skill practice while *self-evaluation* "automatically" improves. We believe that self-evaluation and skill acquisition could be improved at the same time through specific feedback on task skills.

We suggest that specific feedback on task skills should be given when working with any concept map format including CM-c and CM-s. Based on our data, we cannot conclude whether the Kruger-Dunning effect (Kruger and Dunning, 1999) was overcome by training. Nor can we state whether a Kruger-Dunning effect is evident in working with concept maps.

Enjoyment

We hypothesized that CM-c and CM-s trainings increase *enjoyment* during learning with concept maps compared to a control training. Because of missing data, we are unable to answer this research question. Nevertheless, a comparison of CM-c and CM-s learning is legitimate. CM-c and CM-s did not differ in their degree of *enjoyment*. In contrast to Romero et al. (2017), but in line with Blunt and Karpicke (2014), we observed merely moderate *enjoyment* for working with concept maps, while Karpicke and Blunt carried out their study with university students and not school students. We observed in our study higher variability in *enjoyment* than Romero et al. (2017), who carried out their study with medium to high achieving students. Moderate *enjoyment* and high variability in our study, lead us to conclude that concept maps should be applied with the aim to enhance *enjoyment*, especially for those students with yet low to medium academic skills as seen in our study.

Interactive concept maps might provide such an opportunity. Results from meta-analysis have already shown promising effects on learning performance (Schroeder et al., 2018), but the small number of studies does not allow a reliable conclusion. Emotional commitment measured as *enjoyment* is an integral part of meaningful learning. Based on our findings, we recommend taking high variability in *enjoyment* into account and support *enjoyment* for students with the aim to enhance meaningful learning.

Limitations

As common for empirical studies, our results need to be viewed in the context of some limitations. Concerning the measurement of the learning performance, it must be considered that the reliability of the pretest was low ($\alpha = 0.36$). In this study, we intentionally chose a topic that was still unknown to the students of the eighth grade. This guarantees a similar level of prior knowledge. However, it is known that this can lead to a high guessing probability (e.g., Bergman et al., 2015), which in turn can result in poor reliability of the test. Furthermore, we examined *learning performance* immediately after training, as most past findings on trainings on graphic strategies did (Moorf and Readence, 1984). However, delayed learning tests are more sensitive to effects of learning compared to immediate tests (Dunlosky et al., 2013). Future studies might consider analyzing long term effects following concept map trainings to unveil potentially delayed learning effects and we also strongly suggest including motivational measurements as control variables. As most instruments were not designed for the application with junior high school students test validity for this age group has to be confirmed. Moreover, we observed high variability in student's answers, e.g., enjoyment, which reflects "real life" situations but limits options for inferential statistical analyses. Potential effects might be obscured.

Conclusion and practical implications

Acknowledging the limitations of our study, the direct comparison of CM-c and CM-s allows us to contribute to recent meta-analytical findings (Schroeder et al., 2018). In line with Schroeder et al. (2018) we observed that the construction of concept maps has greater impact on cognitive aspects of learning than the study of concept maps. In detail, we found that training in CM-c compared to CM-s training lead to enhanced *learning performance* and *concept map quality*. Concept mapping skills acquired through CM-c training transferred onto learning with CM-s. Students that underwent a CM-c training were able to transfer new skills onto learning with CM-s. We also observed increased accuracy of *self-evaluation* through CM-c training than CM-s training. Beyond these cognitive and metacognitive outcomes, we add insights into emotional effects of learning with concept maps. We found highly dispersed and overall moderate *enjoyment* across students. We did not observe statistically significant differences in *enjoyment* between learning formats after training and learning. Based on the overall results in this study, we conclude that CM-c training has greater effects on cognitive and metacognitive aspects of learning than CM-s training, but not on emotional aspects measured as enjoyment.

For the use in classrooms, we recommend teachers to apply a preceding CM-c training, because it improves *learning performance*, *concept map quality* and students' accuracy of *self-evaluation* compared to CM-s training. Additionally, concept mapping skills acquired through CM-c are likely to be applied by students in learning with CM-c and CM-s similarly. We advise teachers to promote *enjoyment* to enhance long-term commitment with this learning strategy. At the same time, we emphasize high interindividual differences in students' *enjoyment* that needs to be taken into account by teachers. We advise teachers to seek students' direct feedback about *cognitive load* during learning so as to prevent cognitive overload. Concept maps can be applied in many ways and depend on the teacher's goals and the students' needs. This study aimed to contribute to recent knowledge about cognitive, metacognitive and emotional aspects of learning with concept maps, providing aid in choosing suitable learning strategies to support conceptual thinking.

Statements and Declarations

Data Availability Statement

Data are openly available (DOI10.17605/OSF.IO/MW356).

Ethics statement

Ethical review and approval were not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

Author Contributions

SL and JG: conceptualization and methodology. SL and SE: formal analysis, writing – original draft preparation, and visualization. SL: investigation. JG: resources, writing, review, editing and supervision. All authors have read and agreed to the published version of the manuscript.

Funding

This study is part of the project "Learning Biology through Concept Mapping: Importance of a learning strategy training for cognitive load, cognitive processes and learning performance," which was funded by the German Research Foundation (Deutsche Forschungsgemeinschaft; DFG), Grant (GR 4763/2).

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Supplementary Material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/feduc.2022.892312/full#supplementary-material

References

Adesope, O. O., and Nesbit, J. C. (2013). Animated and static concept maps enhance learning from spoken narration. *Learn. Instr. 27*, 1-10. doi: 10.1016/j.learninstruc.2013.02.002

Aguiar, J. G., and Correia, P. R. M. (2017). From representing to modelling knowledge: proposing a two-step training for excellence in concept mapping. *Knowl. Manag. and Elearn Int J. 9*, 366-379. doi: 10.34105/j.kmel.2017.09.022

Amer, A. A. (1994). The effect of knowledge-map and underlining training on the reading comprehension of scientific texts. *Engl. Specif. Purp. 13,* 35-45. doi: 10.1016/0889-4906(94)90023-X

Anderson, J. R. (1983). *The Architecture of Cognition*. Cambridge, MA: Harvard University Press.

Auer, M., Gruber, G., Mayringer, H., and Wimmer, H. (2005). Salzburger *Lese-Screening für die Klassenstufen 5-8* (SLS 5-8) [*Salzburg Reading Screening for Grades 5–8*]. Bern: Huber.

Ausubel, D. P., Novak, J. D., and Hanesian, H. (1978). *Educational psychology - A Cognitive View*. Holt: Rinehart and Winston.

Baumert, J., Lehmann, R., Lehrke, M., Clausen, M., Hosenfeld, I., Neubrand, J., et al. (1998). *Testaufgaben Naturwissenscgaften TIMSS 7./8. Klasse (Population 2) [Test questions natural sciences TIMSS 7./8. Class (population 2)]*. Berlin: Max-Planck-Institut für Bildungsforschung.

Bergman, E. M., de Bruin, A. B., Vorstenbosch, M. A., Kooloos, J. G., Puts, G. C., Leppink, J., et al. (2015). Effects of learning content in context on knowledge acquisition and recall: a pretest-posttest control group design. *BMC Med.l Educ. 15*:133. doi: 10.1186/s12909-015-0416-0

Becker, L. B., Welter, V. D. E., Aschermann, E., and Großschedl, J. (2021). Comprehension-oriented learning of cell biology: do different training conditions affect students' learning success differentially? *Educ. Sci*. 11:438. doi: 10.3390/educsci11080438

Blankenship, J., and Dansereau, D. F. (2000). The effect of animated node-link displays on information recall. *J.Exp.l Educ.* 68, 293-308. doi: 10.1080/00220970009600640

Blunt, J. R., and Karpicke, J. D. (2014). Learning with retrieval-based concept mapping. *J. Educ.l Psychol., 106*, 849-858. doi: 10.1037/a003594

Chang, K.-E., Sung, Y.-T., and Chen, I.-D. (2002). The effect of concept mapping to enhance text comprehension and summarization. *J. Exp. Educ. 71*, 5-23. doi: 10.1080/00220970209602054

Chmielewski, T. C., and Dansereau, D. F. (1998). Enhancing the recall of text: knowledge mapping training promotes implicit transfer. *J. Educ. Psychol. 90*:407. doi: 10.10377/0022-0663.90.3.407

Chularut, P., and DeBacker, T. K. (2004). The influence of concept mapping on achievement, self-regulation, and self-efficacy in students of English as a second language. *Contemp. Educ. Psychol. 29*, 248-263. doi: 10.1016/j.cedpsych.2003.09.001

Clausen, S., and Christian, A. (2012). Concept mapping als messverfahren für den außerschulischen Bereich. [Concept mapping for measurement in a non scholar context]. *J.Didaktik Biowissenschaft, 3*, 18-31.

Collins, A. M., Brown, J. S., and Newman, S. E. (1988). Cognitive apprenticeship: teaching the craft of reading, writing and mathematics. *Think. J. of Philos. Child. 8*, 2-10. doi: 10.5840/thinking19888129

Correia, P. R. M., Infante-Malachias, M. E., and Godoy, C. E. C. (2008). "From theory to practice: the foundations for training students to make collaborative concept maps," in *Proceedings of the 3rd International Conference on Concept Mapping*, Vol. 2, eds A. J. Cañas, J. D. Novak, P. Reiska, and M. K. Ahlberg (Põltsamaa:ValiPress), 414-421.

den Elzen-Rump, V., and Leutner, D. (2007). "Naturwissenschaftliche sachtexte verstehen-ein computerbasiertes trainingsprogramm für schüler der 10. Jahrgangsstufe zum selbstregulierten lernen mit einer mapping-strategie [Understanding scientific factual texts -computer-based training program for 10th grade students for self-regulated learning with a mapping strategy]," in *Selbstregulation Erfolgreich Fördern* [*Successfully Promoting Self-Regulation*], eds. M. Landmann and B. Schmitz (Stuttgart: Kohlhammer), 251-268.

Department of Health, Education and Welfare (2014). The Belmont Report. Ethical principles and guidelines for the protection of human subjects of research. *J. Am. Coll. Dent.* 81:4.

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., and Willingham, D. T. (2013). Improving students' learning with effective learning techniques: promising directions from cognitive and educational psychology. *Psychol.l Sci. Public Interest 14*, 4-58. doi: 10.1177/1529100612453266

Ekinci, S., and Şen, A. İ. (2020). Investigating grade-12 students' cognitive structures about the atomic structure: a content analysis of student concept maps. *Int. J. Sci. Educ. 42*, 977-996. doi: 10.1080/09500693.2020.1744045

Flavell, J. H. (1979). Metacognition and cognitive monitoring: a new area of cognitive–developmental inquiry. *Am. Psychol. 34*:906.

Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., et al. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proc.Natl. Acad. Sci. U.S.A. 111*, 8410-8415. doi: 10.1073/pnas.1319030111

Großschedl, J., and Tröbst, S. (2018). Biologie lernen durch concept mapping: Bedeutung eines Lernstrategietrainings für kognitive Belastung, kognitive Prozesse und Lernleistung - Kurzdarstellung des DFG-projekts [Learning biology by concept mapping: the importance of learning strategy training for cognitive load, cognitive processes and learning performance - brief description of the DFG project]. *Z. Didaktik Biol*. 22, 20–30. doi: 10.4119/zdb-1630

Gehl, D. (2013). Vom Betrachten zum Verstehen *[About Viewing to Understanding]*. Wiesbaden: Springer. doi: 10.1007/978-3-531-19823-1

Hardy, I., and Stadelhofer, B. (2006). Concept Maps wirkungsvoll als strukturierungshilfen einsetzen: welche rolle spielt die selbstkonstruktion? *Z.Pädagog. Psychol. 20*, 175-187. doi: 10.1024/1010-0652.20.3.175

Harmon, M., Smith, T. A., Martin, M. O., Kelly, D. L., Beaton, A. E., Mullis, I. V., et al. (1997). *Performance Assessment: IEA's Third International Mathematics and Science Study (TIMSS).* Amsterdam: International Association for the Evaluation of Educational Achievement.

Hilbert, T. S., Nückles, M., Renkl, A., Minarik, C., Reich, A., and Ruhe, K. (2008). Concept Mapping zum lernen aus texten: können prompts den wissens- und strategieerwerb fördern? [Concept mapping for learning from texts: can prompts promote knowledge and strategy acquisition?]. *Z. pädagog. Psychol. J. Educ. Psychol. 22*, 119-125. doi: 10.1024/1010-0652.22.2.119

Hilbert, T. S., and Renkl, A. (2008). Concept mapping as a follow-up strategy to learning from texts: what characterizes good and poor mappers? *Instr. Sci. 36*, 53-73. doi: 10.1007/s11251-007-9022-9

Holley, C. D., and Dansereau, D. F. (1984). "Networking: the technique and the empirical evidence," In *Spatial Learning Strategies*, eds. D. F. Dansereau and C. D. Holley (Amsterdam: Elsevier), 81-108.

Horton, P. B., McConney, A. A., Gallo, M., Woods, A. L., Senn, G. J., and Hamelin, D. (1993). An investigation of the effectiveness of concept mapping as an instructional tool. *Sci. Educ. 77*, 95-111. doi: 10.1002/sce.3730770107

Hwang, G. J., Chang, S. C., Song, Y., and Hsieh, M. C. (2021). Powering up flipped learning: an online learning environment with a concept map-guided problem-posing strategy. *J.Comput. Assis. Learn., 37*, 429-445. doi: 10.1111/jcal.12499

Ifenthaler, D. (2011). Identifying cross-domain distinguishing features of cognitive structure. *Educ. Technol. Res. Dev. 59*, 817-840. doi: 10.1007/s11423-011-9207-4

Jin, H., and Wong, K. (2010). Training on concept mapping skills in geometry. *J. Math. Educ.* 3, 104-119.

Karpicke, J. D., and Blunt, J. R. (2011). Retrieval practice produces more learning than elaborative studying with concept mapping. *Science*, *331*, 772-775. doi: 10.1126/science.1199327

Keusch, J., and Telaak, S. (2017). *Einfluss der Darstellungsform (Concept Maps vs. Fließtext) auf die Rezeptionsleistung von Schülerinnen und Schülern im Themenbereich „Ökosystem See“ [Influence of the Form of Representation (concept maps vs. continuous text) on the Reception Performance of Schoolchildren in the Subject Area "Lake Ecosystem"]*. Unpublished master's thesis. Köln: University of Cologne.

Klauer, K. J. (1988). Teaching for learning-to-learn: A critical appraisal with some proposals. *Instr.l Sci. 17*, 351-367. doi: 10.1007/BF00056221

Klepsch, M., Schmitz, F., and Seufert, T. (2017). Development and validation of two instruments measuring intrinsic, extraneous, and germane cognitive load. *Front. Psychol. 8*:1997. doi: 10.3389/fpsyg.2017.01997

Klepsch, M., and Seufert, T. (2020). Understanding instructional design effects by differentiated measurement of intrinsic, extraneous, and germane cognitive load. *Instr. Sci. 48*, 45-77. doi: 10.1007/s11251-020-09502-9

Kruger, J., and Dunning, D. (1999). Unskilled and unaware of it: how difficulties in recognizing one’s own incompetence lead to inflated self-assessments. *J.Pers.Soc. Psychol. 77*:1121. doi: 10.1037/0022-3514.77.6.1121

Lambiotte, J. G., Skaggs, L. P., and Dansereau, D. F. (1993). Learning from lectures: effects of knowledge maps and cooperative review strategies. *Appl. Cogn. Psychol. 7*, 483-497. doi: 10.1002/acp.2350070604

Lenhard, W., and Lenhard, A. (2016). *Berechnung von Effektstärken* [*Calculation of Effect Sizes*]. Dettelbach: Psychometrica.

Lenski, S., and Großschedl, J. (2021). *Concept Maps im Unterricht: eine Trainingseinheit für Schüler∗Innen der Sekundarstufe [Concept Maps in the Classroom: a Training Session for Secondary School Students]*. Available online at: https://doi.org/https://doi.org/10.17605/OSF.IO/48A5W

Lenski, S., and Großschedl, J. (im Druck). “Biologie lernen mit concept maps: Lässt sich die Expertise im Umgang mit concept maps von den Augen ablesen? [Learning biology with concept maps: can the expertise in dealing with concept maps be read from the eyes?],” in *Eye Tracking als Methode in der Mathematik- und*

Naturwissenschaftsdidaktik: Forschung und Praxis, eds P. Klein, M. Schindler, N. Graulich, and J. Kuhn (Cham: Springer).

Leopold, C., and Leutner, D. (2015). Improving students' science text comprehension through metacognitive self-regulation when applying learning strategies. *Metacogn. Learn. 10*, 23-27. doi: 10.1007/s11409-014-9130-2

Mayer, R. E., and Moreno, R. (2003). Nine ways to reduce cognitive load in multimedia learning. *Educ. Psychol. 38*, 43-52. doi: 10.1207/S15326985EP3801_6

McCagg, E. C., and Dansereau, D. F. (1991). A convergent paradigm for examining knowledge mapping as a learning strategy. *J. Educ. Res. 84*, 317-324. doi: 10.1080/00220671.1991.9941812

Mintzes, J. J., Canas, A., Coffey, J., Gorman, J., Gurley, L., Hoffman, R., et al. (2011). Comment on "Retrieval practice produces more learning than elaborative studying with concept mapping" [Technical Comment]. *Science,* 334, 453-453. doi: 10.1126/science.1203698

Moorf, D. W., and Readence, J. F. (1984). A quantitative and qualitative review of graphic organizer research. *J. Educ. Res. 78*, 11-17. doi: 10.1080/00220671.1984.10885564

Nesbit, J. C., and Adesope, O. O. (2006). Learning with concept and knowledge maps: a meta-analysis. *Rev. Educ. Res.76*, 413-448. doi: 10.3102/00346543076003413

Nesbit, J. C., and Adesope, O. O. (2011). Learning from animated concept maps with concurrent audio narration. *J. Exp. Educ. 79,* 209-230. doi: 10.1080/00220970903292918

Neuroth, J. (2007). Concept-mapping als Lernstrategie: Eine Interventionsstudie zum Chemielernen aus Texten [Concept mapping as a Learning Strategy: An Intervention Study on Chemistry Learning from Texts]. Berlin:Logos.

North Rhine-Westphalian Ministry of Education Science and Research. (2005). *Schulgesetz für das Land Nordrhein-Westfalen [School law for the state of North Rhine-Westphalia]*. https://recht.nrw.de/lmi/owa/br_vbl_detail_text?anw_nr=6&vd_id=3928&vd_back=N102&sg=&menu=1 (acessed January, 2022).

Novak, J. D., and Cañas, A. J. (2008). *The Theory Underlying Concept Maps And How To Construct Them. Technical Report IHMC CmapTools 2006-01 Rev 01-2008.* FL: Florida Institute for Human and Machine Cognition.

O'Donnell, A. M., Dansereau, D. F., and Hall, R. H. (2002). Knowledge maps as scaffolds for cognitive processing. *Educ. Psychol. Rev.14*, 71-86. doi: 10.1023/A:1013132527007

OECD. (2016). PISA 2015 *Assessment and Analytical Framework: Science, Reading, Mathematic and Financial Literacy*. Paris: OECD Publishing. doi: 10.1787/9789264255425-en

Orru, G., and Longo, L. (2018). “The evolution of cognitive load theory and the measurement of its intrinsic, extraneous and germane loads: a review,” in Proceedings of *International Symposium on Human Mental Workload: Models and Applications*, Vol. 1012, eds L. Longo and M. C. Leva (Cham; Springer), 23-48. doi: 10.1007/978-3-030-14273-5_3

Panadero, E. (2017). A review of self-regulated learning: six models and four directions for research. *Front. Psychol. 8*:422. doi: 10.3389/fpsyg.2017.00422

Patterson, M. E., Dansereau, D. F., and Newbern, D. (1992). Effects of communication aids and strategies on cooperative teaching. *J. Educ. Psychol. 84*, 453–461. doi: 10.1037/0022-0663.84.4.453

Pekrun, R., Elliot, A. J., and Maier, M. A. (2006). Achievement goals and discrete achievement emotions: a theoretical model and prospective test. *J.Educ. Psychol. 98*:583. doi: 10.1037/0022-0663.98.3.583

Poppenk, J., Köhler, S., and Moscovitch, M. (2010). Revisiting the novelty effect: when familiarity, not novelty, enhances memory. *J. Exp. Psychol. Learn., Mem. Cogn. 36*:1321. doi: 10.1037/a0019900

Reader, W., and Hammond, N. (1994). “Computer-based tools to support learning from hypertext: concept mapping tools and beyond,” in *Proceedings of the Computer Assisted Learning: Selected Contributions from the CAL’93 Symposium*, Pergamon, 99-106. doi: 10.1016/B978-0-08-041945-9.50020-2

Renkl, A. (2010). “Lehren und Lernen [Teaching and Learning],” In *Handbuch Bildungsforschung*, eds R. Tippelt and B. Schmidt (Wiesbaden: VS Verlag für Sozialwissenschaften), 737-751. doi: 10.1007/978-3-531-92015-3_39

Renkl, A., and Nückles, M. (2006). “Lernstrategien der externen Visualisierung [External visualization learning strategies],” *Handbuch Lernstrategien*, eds. H. Mandl. And H. F. Freidrich (Göttingen: Hogrefe), 135-147.

Rewey, K. L., Dansereau, D. F., Skaggs, L. P., and Hall, R. H. (1989). Effects of scripted cooperation and knowledge maps on the processing of technical material. *J. Educ. Psychol. 81*, 604–609. doi: 10.1037/0022-0663.81.4.604

Romero, M. d. C., Cazorla, M., and Buzón García, O. (2017). Meaningful learning using concept maps as a learning strategy. *J.Technol. Sci. Educ.* 7, 313-332. doi: 10.3926/jotse.276

Royer, J. M. (1979). Theories of the transfer of learning. *Educ. Psychol. 14*, 53-69. doi: 10.1080/00461527909529207

Royer, J. M., and Cable, G. W. (1976). Illustrations, analogies, and facilitative transfer in prose learning. *J.Educ. Psychol. 68,* 205. doi: 10.1037/0022-0663.68.2.205

Ruiz-Primo, M. A. (2004). “Examining concept maps as an assessment tool,” in: *Proceedings of the 1st International Conference on Concept Mapping*, Pamplona.

Salata, M. W. A. (1999). *Concept Maps as Organizers in an Introductory University Level Biology Course*. Charlottesville, VA: University of Virginia.

Schraw, G. (1998). Promoting general metacognitive awareness. *Instr. Sci. 26*, 113-125. doi: 10.1023/A:1003044231033

Schroeder, N. L., Nesbit, J. C., Anguiano, C. J., and Adesope, O. O. (2018). Studying and constructing concept maps: a meta-analysis. *Educ. Psychol. Rev. 30*, 431-455. doi: 10.1007/s10648-017-9403-9

Sumfleth, E., Neuroth, J., and Leutner, D. (2010). Concept mapping – eine lernstrategie muss man lernen. [Concept mapping – learning strategy is something you must learn]. *Chemkon 17,* 66-70. doi: 10.1002/ckon.201010114

Sweller, J. (2010). Element interactivity and intrinsic, extraneous, and germane cognitive load. *Educ. Psychol. Rev. 22,* 123-138. doi: 10.1007/s10648-010-9128-5

Tseng, S.-S. (2020). Using concept mapping activities to enhance students’ critical thinking skills at a high school in taiwan. *Asia Pac. Educ. Res. 29*, 249-256. doi: 10.1007/s40299-019-00474-0

Visible learning Meta^X research base® (2021). *Global research Data Base*. *CORWIN. Version 1.1. updated August 2021.* Available online at: https://www.visiblelearningmetax.com/influences (accessed December 21, 2021).

Wild, K.-P. (2001). “Lernstrategien und Lernstile [learning strategies and learning styles],” in *Handwörterbuch Pädagogische Psychologie [Handbook of educational psychology]*, ed. D. H. Rost (Weinheim: Beltz, Psychologie VerlagsUnion), 309-312.

Woldeamanuel, Y. W., Abate, N. T., and Berhane, D. E. (2020). Effectiveness of concept mapping based teaching methods on grade eight students’ conceptual understanding of photosynthesis at ewket fana primary school, Bahir Dar, Ethiopia. *EURASIA J. Math. Sci. Technol. Educ.* 16, 1-16. doi: 10.29333/ejmste/9276

Young, J. Q., Van Merrienboer, J., Durning, S., and Ten Cate, O. (2014). Cognitive load theory: implications for medical education: AMEE Guide No. 86. *Med. Teach.* 36, 371-384. doi: 10.3109/0142159X.2014.889290

Zimmerman, B. J. (2000). “Attaining self-regulation: A social cognitive perspective,” in *Handbook of Self-Regulation*, eds M. Boekarts, P. R. Pintrich, and M. Zeidner (San Diego, CA: Academic Press), 13-39. doi: 10.1016/B978-012109890-2/50031-7

3.2 Study II: Promoting Self-Evaluation through Prompting

Study II was published in 2024 in the *European Journal of Psychology of Education*. The study can be found under the following citation:

> Elsner, S. & Großschedl, J. (2024). Can metacognitive accuracy be altered through prompting in biology text reading?. *European Journal of Psychology of Education*, *39*(2), 1465-1483. DOI: https://doi.org/10.1007/s10212-023-00747-9

3.2.1 Study II: Summary, Research Questions and Methodological Approach

Study II examines whether the accuracy of self-evaluation can be altered through specific instruction during the learning process. The cue-utilization framework suggests that self-evaluations are formed based on incoming information and it’s processing (Koriat, 1997). However, not all incoming information is predictive of actual performance, and the use of less predictive cues is associated with lower accuracy of self-evaluation, e.g., Thiede et al. (2010). The embedded model of working memory (Cowan, 1988) may help explain how prompts could increase the accuracy of self-evaluation, as prompts may direct learners’ attention from less predictive to more predictive cues. In the present study, it is hypothesised that the use of incoming information can be guided through prompts. Depending on the type of prompt self-evaluation may be positively or negatively affected. The effects of resource-oriented and deficit-oriented prompts are examined in Study II. This study addresses the following research questions:

> **RQ 1:** Does prompting during text reading alter the accuracy of self-evaluation of text comprehension?
>
> **RQ 2:** How do resource-oriented and deficit-oriented prompts influence accuracy of self-evaluation?
>
> **RQ 3:** Does prompting enhance learning performance?

To address these research questions, an online study was conducted with 162 pre-service biology teachers. Participants were asked to read a biology text, assess their level of comprehension, and answer test questions to “objectively” measure their text comprehension. Participants were randomly assigned to one of three conditions: a resource-oriented question (i.e., “What have I already understood?”), a deficit-oriented question (i.e., “What have I not yet understood?”), or no question at all. To determine whether the deliberate instruction affected self-evaluation, the discrepancy between self-evaluation and “objectively" measured learning performance was calculated. For this, learning performance was subtracted from the judgment of comprehension, resulting in negative values indicating underestimation and positive values indicating overestimation. The discrepancy values were compared across the three groups. The limitations of this method in determining the accuracy of self-evaluation will be discussed in Chapter 4.2 of this dissertation.

3.2.2 Study II: Own Contribution

The author of this dissertation designed and planned this study. The author was responsible for data acquisition, data preparation, data analysis, and writing. The author also prepared the data for open-access publication through the Open Science Framework (https://osf.io/ykzvg/). The authors’ contributions can also be found in the original manuscript.

3.2.3 Study II: Published Manuscript

Can metacognitive Accuracy be altered through Prompting in Biology Text Reading?

Elsner, S. and Großschedl, J.

Abstract

Metacognitive accuracy is understood as the congruency of subjective evaluation and objectively measured learning performance. With reference to the *cue utilisation framework* and the *embedded-processes model of working memory*, we proposed that prompts impact attentional processes during learning. Through guided prompting, learners place their attention on specific information during the learning process. We assumed that the information will be taken into account when comprehension judgments are formed. Subsequently, metacognitive accuracy will be altered. Based on the results of this online-study with pre-service biology teachers, we can neither confirm nor reject our main hypothesis and assume small effects of prompting on metacognitive accuracy if there are any. Learning performance and judgment of comprehension were not found to be impacted by the use of resource- and deficit-oriented prompting. Other measurements of self-evaluation (i.e., satisfaction with learning outcome and prediction about prolonged comprehension) were not influenced through prompting. The study provides merely tentative evidence for altered metacognitive accuracy and effects on information processing through prompting. Results are discussed in light of online learning settings in which the effectiveness of prompt implementation might have been restricted compared to a classroom environment. We provide recommendations for the use of prompts in learning settings with the aim to facilitate their effectiveness, so that both resource-oriented and deficit-oriented prompts can contribute to metacognitive skill development if they are applied appropriately.

Keywords: metacognition, accuracy, resource-orientation, deficit-orientation, prompts

Background

Metacognitive accuracy and metacognition

Our modern and rapidly changing world demands flexible and fast learning in many areas of life (e.g., grasping new software features after changing a job or developing communication skills in order to optimise work efficiency). The abilities to set learning goals, to watch our progress and to assess goal attainment are vital to adapt to new challenges. Evolving the ability to evaluate one's own learning process accurately is inherent to learners' skill development towards lifelong self-regulated learning (e.g., de Boer et al., 2018). The congruency of subjective evaluation of one's own learning and objectively measured learning performance can be defined as metacognitive accuracy. It is located within the evaluating domain of metacognitive processes that have been investigated since the introduction of the term 'metacognition' into educational research and practice in the 70s (Flavell, 1979). Generally, metacognition refers to three domains of learning: planning, monitoring and evaluating (Schraw & Moshman, 1995). These domains incorporate metacognitive knowledge (i.e., knowledge about person variables, task features and learning strategies) and active regulation of cognition (i.e., skills and processes that guide, monitor, control and regulate learning; Veenman, 2012).

The importance of metacognitive accuracy and the impact of instructional practices

Metacognitive evaluations and their accuracy can drive future learning behaviour and its continuity. This has been observed in experimental studies (e.g., Mazzoni & Cornoldi, 1993; Mazzoni et al., 1990, Metcalfe & Finn, 2008; Mitchum et al., 2016; Rhodes & Castel, 2009) and suggested by theoretical approaches like the 'region of proximal learning model of study time allocation' (Metcalfe & Kornell, 2005) and 'discrepancy-reduction' models (summarised in Thiede et al., 2003). Metacognitive accuracy has been shown to influence regulation of learning and learning performance (Thiede et al., 2003). At the same time, metacognitive evaluations are prone to errors (Dunlosky & Lipko, 2007; Dunning et al., 2004) and metacognitive accuracy might be comparably low without specific instructional practices aiming to improve accuracy. For example, summarising, re-reading, retrieval practice or delaying time between study phase and metacognitive evaluation (*Delayed Judgment of Learning Effect)* lead to an increase in metacognitive accuracy (Miller & Geraci, 2014; Nelson & Dunlosky, 1991; Rawson et al., 2000; Thiede et al., 2003; Thiede et al., 2005). In detail, delaying judgment intervals and retention time, matching judgment of learning items with test questions, and applying cued recall tasks in learning tests are associated with higher metacognitive accuracy (Rhodes & Tauber, 2011).

Given the far-reaching consequences of inaccurate metacognitive evaluations (e.g., stopping study efforts without realising that expectations have not yet been met due to overestimation), continuous investigation of influencing factors is needed (for a prior summary, see Thiede et al., 2003). Research about underlying cognitive processes through specific instructional practices contributes to a better understanding of the formation of metacognitive evaluations, their accuracy and the link to adaptive and effective learning behaviour.

Metacognitive evaluation and information processing via cue utilisation

According to the *cue-utilisation framework* (Koriat, 1997), the formation of metacognitive evaluations relies on 'incoming' information ('cues') during learning instead of memory traces being directly utilised as proposed by King et al. (1980). Such cues might include task-specific, content-specific, emotion-related, and behavioural-related information. Other information might be drawn from past experiences or expectations about the future (see Table 1 for examples). Because of the large number of cues that learners might focus on - even on multiple cues simultaneously (Undorf et al., 2018) - it seems conclusive that some cues are more predictive of future performance than others. Indeed, it was shown that comprehension-based cues, such as the self-judged ability to explain a text, are more predictive of performance than information about the quality of a text itself (Thiede et al., 2010). It is assumed that the use of less predictive cues relates to lower metacognitive accuracy (Prinz-Weiß et al., 2022; Serra & Dunlosky, 2010; Thiede et al., 2010). Nevertheless, learners' attention might be directed towards more predictive cues through instructional practices (e.g., use of prompts). We assume that such deliberate and specific guidance of information processing impacts metacognitive accuracy via the allocation of attention.

Table 1

Classification of Possible Information Integrated into Metacognitive Evaluations

Type of information	Examples
Task-specific information	Does the task have specific requirements (e.g., drawing, discussing, reading, etc.)? What kind of retrieval is expected (e.g., word recall or word recognition)?
Content-specific information	How much did I know about this topic beforehand? Does the topic cover basic or complex matters?
Emotion-related information	How did I feel during studying? Did it feel easy or difficult to read the text? Did I enjoy reading?
Behavioural-related information	Did I have to look up technical terms? Was I able to concentrate and stick to the task?
Information about past experiences	How well did I do on past tests in general? How well did I do on past tests on this topic?
Information about expectations	Will I be able to explain the content to others later?
Comparative information	How well do others typically do? Did I perform better than others?

Prompts and their link to performance outcome and metacognitive accuracy

Prompts are typically applied by teachers to guide information processing and scaffold students learning. Metacognitive prompts can take the shape of questions or cues and target learners' monitoring abilities. Metacognitive prompts are the most widely studied practice of metacognitive instructions (Zohar & Barzilai, 2013). They can be combined with cognitive instructions (e.g., Berthold et al., 2007; Hübner et al., 2006) or broader instructional approaches such as context-based learning (e.g., Dori et al., 2018). Recent research shows a special interest in computer-based learning (Bannert & Mengelkamp, 2013; Bannert & Reimann, 2012; Daumiller & Dresel, 2019; Van den Boom, 2004; Zheng, 2016) and self-directed metacognitive prompts (Bannert et al., 2015; Engelmann et al., 2021). Single studies show that metacognitive prompts aligned with cognitive tasks improved learning outcome in psychology students (Berthold et al., 2007) and biology students (Großschedl & Harms, 2013). Metacognitive prompts in context-based learning improved scientific understanding in chemistry students (Dori et al., 2018). Metacognitive prompts also increased understanding of the nature of science in pupils (Peters & Kitsantas, 2010). Some studies did not observe effects on learning outcome (e.g., McCarthy et al., 2018; Moser et al., 2017; van Alten et al., 2020), but on task completion rate (van Alten et al., 2020) and qualitative reports of goal setting (McCarthy et al., 2018). Metacognitive processes (i.e., self-awareness) were improved through generic prompts during learning (Kramarski & Kohen, 2017). One study emphasises that metacognitive prompts are only effective when they are used regularly and with elaborate note-taking (Moser et al., 2017). In general, an increasing number of studies provide evidence for enhanced learning performance and increased metacognitive activity through metacognitive prompts (Devolder et al., 2012; Donker et al., 2014; Haller et al., 1988; Kim et al., 2018; Zohar & Barzilai, 2013).

Less is known about the effectiveness of prompts on metacognitive accuracy, and findings are inconclusive. For example, cognitive and metacognitive prompts were not found to increase metacognitive accuracy (Berthold et al., 2007). At the same time accuracy ranged widely in different prompting conditions in this study. A more recent study suggests alterations in cue-use through variation of prompt frequency (Vangsness & Young, 2021). However, questions about the link to metacognitive accuracy remain unanswered. Potential effects of prompting on metacognitive accuracy might be derived from the *embedded-processes model of working memory* (Cowan, 1988).

Information processing via cue utilisation, prompts and attention allocation

The *embedded-processes model of working memory* (Cowan, 1988) proposes that information need to obtain a state of availability in order to be utilised for the execution of a task (here: formation of metacognitive evaluations). Information might reach different, hierarchically structured states of availability – from long-term memory (a), and an activated state of long-term memory (b) to a 'focus of attention' (c). At the highest level of the hierarchy, information is highly accessible if it reaches the state of 'focus of attention'. Attention can be controlled by voluntary (central executive function) and involuntary processes (attentional orienting system). We expect that prompts impact these attentional processes.

We assume that prompts initially provide a stimulus to the attentional orienting system: they are able to direct learners' attention to specific cues during learning. Ideally, attention is drawn to cues that are predictive of performance. Simultaneously, prompts stimulate the central executive and 'encourage' regulation of attention. Information processing shifts from bottom-up to top-down regulation. In this way, prompts stimulate metacognitive activity and actions of self-regulated learning. Which information is made available for the subsequent formation of metacognitive evaluation is determined by the nature of a prompt. Resource-oriented and deficit-oriented prompts can be derived from the field of developmental psychology (Petermann & Schmidt, 2006). Both types of prompts might be applied during learning, i.e., 'What have I already understood' vs. 'What have I not yet understood?'. Both prompts are comprehension-based, non-specific and versatile in their application (for practical use of these questions see: Schraw, 1998). If these prompts are applied during text-reading, we assume attention to be discriminatively allocated.

If the resource-oriented prompt is applied, the focus of attention is directed towards one's own comprehension (i.e., content of the topic that is already known to the learner). This includes currently acquired knowledge and also prior knowledge acquired through previous learning opportunities. Although comprehension-based information is likely to recede from the 'focus of attention' after reading, they will remain in an increased state of availability (activated state of working memory) for subsequent evaluations and serve as information that improve metacognitive accuracy. Because we expect learners to internally repeat topic content, the effects of resource-oriented prompts are potentially comparable to the effects of retrieval practice (Miller & Geraci, 2014). Resource-oriented prompts supposedly lead to similar improvements in metacognitive accuracy.

If the deficit-oriented prompt is applied, information processes might be altered in multiple ways. First, attention might initially be directed towards one's own comprehension as a benchmark measure in order to identify what has not yet been understood as suggested by *discrepancy-reduction models* (see: Thiede et al., 2003). The effects are likely to be similar compared to those following the use of resource-oriented prompts. Second, applying the deficit-oriented prompt might direct learners' attention away from internal comprehension-based information towards text passages that have not yet been understood (external information). Immediate regulation of learning behaviour takes place and learners focus their attention on 'new' content. This might increase the total amount of available comprehension-based cues and may enhance metacognitive accuracy beyond the effects of resource-oriented prompts. Effects on metacognitive accuracy might be similar to those of re-reading methods (Rawson et al., 2000). While these first two mechanisms might occur in parallel and increase metacognitive accuracy, a third mechanism might be detrimental to metacognitive accuracy. Instead of a redirection towards well-understood content or passages that are yet to be understood, attention might be directed to a lack of understanding (void of comprehension) but because this lack of understanding itself is not a valid information on which attention could be placed on, information processing will be interrupted (or even stopped). Learning behaviour is not regulated and metacognitive accuracy will either be unaffected or negatively affected.

Study aim and hypotheses

This study aims to extend current scientific evidence about instructional practices that are applied to increase metacognitive accuracy. We investigate the effects of two comprehension-based and non-specific prompts (i.e., resource-oriented and deficit-oriented) on the congruency of subjective evaluation and objectively measured text comprehension after reading a biology text. This study tests the hypothesis that prompts direct learners' attention and impact metacognitive accuracy. In detail, we hypothesise that applying resource-oriented prompts leads to an increase in metacognitive accuracy, and we propose multiple mechanisms when applying a deficit-oriented prompt.

Methods

Participants

A total of 162 pre-service biology teachers took part in this study. On average, university students were 25.18 years old (SD = 3.62 years) and 80.2 % were female, 17.9 % were male, 0.6 % were non-binary, and 1.2 % made no gender specification. At the time of the assessment students studied biology education for various school forms in Germany: vocational college ('Berufsschule', *n* = 2) elementary school ('Grundschule', *n* = 7), non-academic track secondary school ('Hauptschule, Realschule, Sekundarstufe, Gesamtschule', *n* = 45), academic track grammar school ('Gymnasium/ Gesamtschule', *n* = 61) and special needs education ('Sonderpädagogische Förderung', *n* = 47). Eleven students attended the bachelor's programme in biology teaching, 151 attended the equivalent master's programme.

General study design

This study was designed as an online learning experiment and distributed via a survey link. The online survey was designed with SoSci Survey (Leiner, 2019). The experiment included four parts: gathering of demographic information [1], text-reading in either one of three conditions [2], self-evaluation of the learning process [3], and the measurement of learning performance [4].

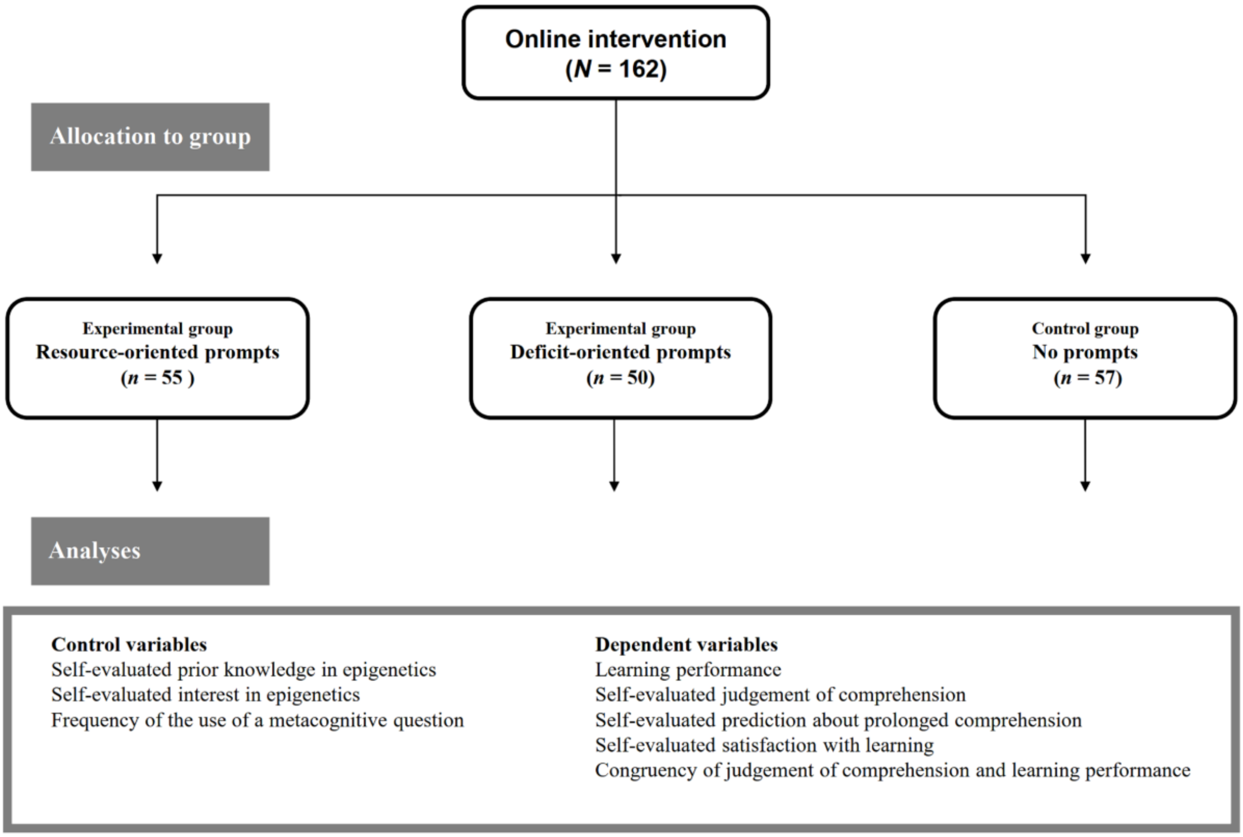
In part [2] – text-reading, participants were randomly assigned to one of three groups i.e., two experimental groups that were prompted to use a metacognitive question and one control group without the use of prompts (see figure 1 for participant allocation). In one experimental group, participants were prompted to apply the resource-oriented metacognitive question ‘What have I already understood?’. In the other experimental group, participants were prompted to apply the deficit-oriented metacognitive question ‘What have I not yet understood?’. Prompts were placed at four positions throughout the text: at the beginning, after the first and the second third of the text, and at the end of the text. Participants in both experimental groups were prompted to apply and answer the respective metacognitive question for themselves. Fifty-five participants were prompted to use a resource-oriented metacognitive question, 50 participants were prompted to use a deficit-oriented metacognitive question, and 57 participants were not prompted to use a metacognitive question.

Procedure

For this study, pre-service biology teachers were recruited from a university course in biology education in Germany. Pre-service teachers were invited to take part in this online learning experiment via email. They were also asked to share the survey link with any biology pre-service teacher interested in participating in the study. Participants were informed about the general research aim, the study procedure including study length and a planned learning test, criteria of eligibility, voluntariness and the non-risk character of the study. Protection of data privacy was ensured and participants were informed about the possibility to withdraw from participation at any time. Participants gave their consent after reading the study information by clicking ‘continue’. We gathered demographic information including age, gender, school form, and second subject. Participants were then randomly assigned to one of three groups (i.e., two experimental groups that were prompted to use a metacognitive question and one control group without the application of prompts). In accordance with recommendations for metacognition teaching (Schraw, 1998), a short written introduction about metacognition in learning settings was given to both experimental groups to inform them about the usefulness of metacognitive learning strategies. Participants in the control group received neither an introduction about the usefulness of metacognitive learning strategies nor any prompts to apply a metacognitive question. Subsequently, participants were asked to read a book chapter of approximately 1500 words retrieved from a teaching book about ‘epigenetics’ (Knippers, 2017). The chapter discussed the origin of epigenetics, DNA methylation and differences between identical genes. We expected a reading duration of approximately fifteen minutes. To ensure thorough reading, participants were informed that they can continue to the next survey page when they have spent at least 10 minutes on the text-containing page. After text-reading, all participants were asked to self-evaluate their learning process and to take a short learning test about the text content (i.e., epigenetics). After data submission, participants were shown correct answers to the learning test question.

Figure 1

Overview of Particpant Allocation, Control Variables and Dependent Variables in this Study


Online intervention
(*N* = 162)
Allocation to group
Experimental group
Resource-oriented prompts
(*n* = 55)
Experimental group
Deficit-oriented prompts
(*n* = 50)
Control group
No prompts
(*n* = 57)
Analyses
Control variables
Self-evaluated prior knowledge in epigenetics
Self-evaluated interest in epigenetics
Frequency of the use of a metacognitive question
Dependent variables
Learning performance
Self-evaluated judgement of comprehension
Self-evaluated prediction about prolonged comprehension
Self-evaluated satisfaction with learning
Congruency of judgement of comprehension and learning performance

Instruments

Self-evaluation of the learning process

Single self-report questions were applied to measure self-evaluation in different manifestations. Prior knowledge about epigenetics and interest in the topic were gathered. Beyond that, frequency of the use of a metacognitive question, satisfaction with the learning outcome and students' judgment of comprehension were measured. Students were also asked to make a prediction of prolonged comprehension. Prior knowledge about epigenetics was evaluated on a visual analogue scale from 0% to 100% ('How much of the text have you known before reading?'). Interest in epigenetics and satisfaction with learning outcome were rated on a visual analogue scale from 'not at all' (= 0) to 'completely' (= 100) through agreement with the statements 'I find the topic epigenetics interesting' and 'I am satisfied with my learning outcome'. The use frequency of a metacognitive question was captured by the question 'How often did you apply the/a metacognitive question?' on an eight-stepped scale ('never', 'one time', 'two times', 'three times', 'four times', 'five times', 'six times', 'more than six times'). The control group received a one-sentence explanation about the meaning of metacognitive questions to account for a potential lack of knowledge about metacognitive learning strategies. Judgment of comprehension and a prediction about prolonged comprehension were made on a visual analogue scale from 0% to 100% ('How much of the content did you comprehend?' and 'How much of the content will you still know in one week?').

Learning performance

Text comprehension was measured with a learning test about epigenetics. This learning test was designed based on the chapter that students were asked to read (Knippers, 2017). It consisted of seven closed, single-choice questions and eight open-ended questions (see supplementary material for learning test questions and assessment criteria). Sample questions are 'Early observations suggest a relationship between methylation of cytosine bases and the gene regulation. Describe these observations!' and 'Which answer is correct? Patterns of methylation …' with the response options '…vary from cell to cell', 'vary from person to person', '…can alter during the course of a life' and 'all of these answers are correct'. One point was assigned to each correct closed single-choice question. Two points were assigned to correct open-ended questions. One point was assigned to partially correct open-ended questions. Incorrect answers received no point. All open-ended questions were rated by one rater. Additionally, two staff members in research positions rated twenty percent of the material. Based on mean ratings ($k = 3$), intraclass correlation coefficients were calculated using a one-way, random model in the SPSS version 28.0. Intraclass estimates revealed moderate to excellent agreement in reference to (Koo & Li, 2016) between the three raters across all eight open-ended questions (.66 to .96). Learning performance was measured as a sum score that was transformed into a percentage value between 0% and 100%. As a measurement of reliability, we report Cronbach's α of .79 across all 15 items of the learning test.

Data preparation and data analyses

Prior to data acquisition, power analyses were carried out to determine the optimal number of participants. A priori power analyses for a one-way ANOVA with three groups, an expected effect size of $f = 0.25$, α-level of .05 and a favoured power of .80, resulted in a total sample size of 159 participants. Actual sample size reached 162 participants.

As a measurement of metacognitive accuracy, we subtracted learning performance from judgment of comprehension. It reflects the discrepancy between subjective evaluation of one's own learning and objectively measured learning performance. A value of zero indicates complete congruity. A negative value indicates an underestimation. A positive value indicates an overestimation.

Prior to data analyses, all data were checked for extreme values. Values that exceeded the threshold of two standard deviations from the group mean were excluded from analyses (Simmons et al., 2011). For inferential statistical analyses, respective assumptions were tested. In case of a violation, alternative tests were applied and are being reported where they are applied. In accordance with recommendations by Döring & Bortz (2016, pp. 673 – 674) and well-known criticism about the use and interpretation of *p*-values (e.g., Gardner & Altman, 1986; Wasserstein & Lazar, 2016), we report 95% confidence intervals for mean values and effects sizes in addition to typical *p*-value interpretation. For inferential statistics in which the null hypothesis was 'favoured', we carried out analyses at an α-level of 0.10 because increasing the α-level allows to indirectly minimise the β-error in statistical analyses (Döring & Bortz, 2016, pp. 885 – 888). We adjusted *p*-values for multiple testing in all analyses in which the alternative hypothesis was favoured. We adjusted according to the Bonferroni-Holm method (Hemmerich, 2016; Holm, 1979). Original and adjusted values are reported where they were applied. Most statistical analyses were carried out with IBM SPSS statistics, version 28.0. If not provided by SPSS, effect sizes were calculated in https://www.psychometrica.de. Data are openly available in DOI 10.17605/OSF.IO/YKZVG.

Results

Metacognitive accuracy - Congruency of judgment of comprehension and learning performance

This study aimed to investigate potential effects of resource-oriented and deficit-oriented prompting on pre-service teachers' metacognitive accuracy after text-reading. It reflects the discrepancy between subjective evaluation of one's own learning and objectively measured learning performance. A value of zero indicates complete congruity. A negative value indicates an underestimation. A positive value indicates an overestimation (see Table 2 for descriptive data and inferential statistics). Our results show overall positive means indicating overestimation in metacognitive accuracy in all three groups (resource-oriented prompting: *M* = 18.24, 95%, *SD* = 21.85, *n* = 55; deficit-oriented prompting: *M* = 28.40, 95%, *SD* = 18.64, *n* = 47; no prompting M = 18.54, *SD* = 24.14, *n* = 52). We compared pre-service teachers' metacognitive accuracy between the three groups. After adjusting for multiple testing, we did not observe statistically detectable differences between resource-oriented prompting, deficit-oriented prompting and no prompting applying a Welch ANOVA with heterogeneity of variances *F*(2, 100.17) = 4.07; *p* = .020, $p_{adj.}$ = .140; η^2 = .044. We also investigated the effect of the use frequency of the metacognitive question as an indicator of instruction efficacy. A rank analysis of covariance (Quade, 1967) with use frequency as covariate yielded similar results (*F*(2,100.17) = 4.04; *p* = .021, p_{adj}. = .140) as the Welch ANOVA, unexpectedly suggesting no meaningful impact of use frequency of a metacognitive question.

Nevertheless, 95% confidence intervals provide indication for an increased overestimation elicited through deficit-oriented prompting. Confidence intervals for mean metacognitive accuracy after resource-oriented prompting and no prompting largely overlap. We observed a confidence interval of 12.34 to 24.15 after resource-oriented prompting and a confidence interval of 11.82 to 25.26 after no prompting. The confidence interval of 22.93 to 33.87 for mean accuracy after deficit-oriented prompting is somewhat shifted towards positive values. Effect sizes are small to medium. Based on *p*-values, confidence intervals for mean values and effect sizes, we can neither confirm nor reject our hypothesis regarding the effects of prompting on metacognitive accuracy.

We neither observed statistically detectable differences in learning performance applying a Kruskal-Wallis test $\chi^2(2)$ = 0.21, *p* = .901; $p_{adj.}$ > .999; η^2 = .011 or in judgment of comprehension applying a Kruskal-Wallis test $\chi^2(2)$ = 3.39, *p* = .184; $p_{adj.}$ > .920; η^2 = .009. Overall, our results revealed mean learning performance of 44.1% (*SD* = 20.84 %) and mean judgment of comprehension of 65.5% (*SD* = 21.62%).

Table 2

Descriptive and inferential data for all dependent variables

	Group	*M*	95% *CI*	*SD*	*n*	Inferential analysis	Effect size
Metacognitive Accuracy in %	Resource-oriented prompt	18.24	12.34 - 24.15	21.85	55	Welch ANOVA	
	Deficit-oriented prompt	28.40	22.93 - 33.87	18.64	47	$F(2, 100.17) = 4.07$	$\eta^2 = .044$
	No prompt	18.54	11.82 - 25.26	24.14	52	$p = .020$, $p_{adj.} = .120$	
Learning Performance in %	Resource-oriented prompt	44.28	38.60 - 49,97	20.83	54	Kruskal-Wallis test	
	Deficit-oriented prompt	42.43	37.06 - 47.81	18.91	50	$\chi^2(2) = 0.21$, $p = .901$	$\eta^2 = .011$
	No prompt	44.56	37.58 - 49.53	22.30	56	$p_{adj.} > .999$	
Judgment of Comprehension in %	Resource-oriented prompt	68.48	63.88 - 73.08	16.19	50	Kruskal-Wallis test	
	Deficit-oriented prompt	72.78	68.71 - 76.86	13.71	46	$\chi^2(2) = 3.39$, $p = .184$	$\eta^2 = .009$
	No prompt	64.38	58.32 - 70.45	22.43	55	$p_{adj.} > .920$	
Satisfaction with learning Outcome in %	Resource-oriented prompt	63.20	58.32 - 68.08	17.88	54	Kruskal-Wallis test	
	Deficit-oriented prompt	60.29	54.60 - 65.98	19.81	49	$\chi^2(2) = 0.87$, $p = .648$	$\eta^2 = .007$
	No prompt	59.09	53.28 - 64.91	21.11	53	$p_{adj.} > .999$	
Prediction about prolonged Comprehension in %	Resource-oriented prompt	36.86	31.93 - 41.76	17.65	52	Kruskal-Wallis test	
	Deficit-oriented prompt	39.94	33.94 - 45.93	20.65	48	$\chi^2(2) = 2.87$, $p = .238$	$\eta^2 = .006$
	No prompt	32.96	27.53 - 38.40	20.47	57	$p_{adj.} > .952$	
Prior knowledge	Resource-oriented prompt	31.89	26.44 - 37.33	19.75	53	ANOVA	
	Deficit-oriented prompt	39.52	33.55 - 45.49	20.99	50	$F(2, 156) = 1.87$	$\eta^2 = .023$
	No prompt	35.21	29.99 - 40.44	19.52	56	$p = .158$	
Interest	Resource-oriented prompt	76.47	70.75 - 82.19	20.76	53	Kruskal-Wallis test	
	Deficit-oriented prompt	75.73	69.48 - 81.99	21.79	49	$\chi^2\ (2) = 0.63$	$\eta^2 = .009$
	No prompt	73.63	67.52 - 79.75	21.29	49	$p = .728$	

Self-evaluated satisfaction with learning outcome and prediction about prolonged comprehension

We tested whether resource-oriented prompting, deficit-oriented prompting, and no prompting influenced satisfaction with the learning outcome and prediction about prolonged comprehension (see Table 2 for descriptive data) via Kruskal-Wallis tests. We did not observe a statistically detectable difference in satisfaction with learning outcome between resource-oriented, deficit-oriented prompting and no prompting; $\chi^2(2) = 0.87$, $p = .648$; $p_{adj.} > .999$; $\eta^2 = .007$. We did not observe a statistically detectable difference in the prediction about prolonged comprehension ('How much of the content will you still know in one week?')¸ $\chi^2(2) = 2.87$, $p = .238$; $p_{adj.} > .952$; $\eta^2 = .006$.

Preliminary tests

Preliminary analyses were carried out to ensure absence of substantial differences between the three groups (resource-oriented prompting, deficit-oriented prompting, no prompting) at baseline (see table 2 for descriptive data). Preliminary tests revealed no statistically detectable differences in age; $F(2, 159) = 0.56$, $p = .571$, or prior knowledge; $F(2, 156) = 1.87$; $p = .158$; $\eta^2 = .023$ between groups. We observed ratings of prior knowledge at a moderate level with high variance across all groups; resource-oriented prompting: $M = 31.89$, 95% *CI* [26.44; 37.33], $SD = 19.75$, $n = 53$, deficit-oriented prompting: $M = 39.52$, 95% *CI* [33.55; 45.49], $SD = 20.99$, $n = 50$, no prompting: 35.21, 95% *CI* [29.99; 40.44], $SD = 19.52$, $n = 56$. Interest in the topic 'epigenetics' did not differ between the groups; χ^2 (2) = 0.63, $p = .728$; $\eta^2 = .009$. We observed rather high mean ratings of interest in the topic 'epigenetics' with high variance across all groups; resource-oriented prompting: $M = 76.47$, 95% *CI* [70.75; 82.19], $SD = 20.76$; $n = 53$, deficit-oriented prompting: $M = 75.73$, 95% *CI* [69.48; 81.99], $SD = 21.79$; $n = 49$, no prompting: $M = 73.63$, 95% *CI* [67.52; 79.75], $SD = 21.29$, $n = 49$.

Manipulation check

To ensure that prompting did indeed increase the use of a metacognitive question as intended through prompting, we ran a Kruskal-Wallis-Test with ordinal scaled data for use frequency of a metacognitive question. Unexpectedly, no statistically detectable difference between the three groups was observed ($\chi^2(2) = 1.95$, $p = .378$; $p_{\text{adj.}} > .999$) with mean ranks for use frequency of 84.85 (resource-oriented prompt), 85.53 (deficit-oriented prompt) and 74.74 (no prompt). Using a metacognitive question three times was most frequently reported in the group with the resource-oriented question (23 times, 42%) and the deficit-oriented question

(15 times, 30%). Using a metacognitive question two times was most frequently reported in the control group (18 times, 32%). This unexpected outcome will be discussed in the limitations section.

Discussion

How accurate are metacognitive judgments of comprehension?

The main aim of this study was to examine the impact of resource-oriented and deficit-oriented prompts during text-reading on metacognitive accuracy when evaluating text comprehension. Our hypothesis was based on the *embedded-processes model of working memory* (Cowan, 1988), and the notion that metacognitive prompts can be used to allocate attention to specific features during learning. We assumed that the information will subsequently be used when forming a judgment about comprehension. We hypothesised that applying resource-oriented prompts leads to an increase in metacognitive accuracy, and we propose multiple mechanisms when applying a deficit-oriented prompt. To test this hypothesis we examined the discrepancy of judgment of comprehension and learning performance between groups.

Based on our analyses, we can neither confirm nor reject our hypothesis. We observed shifted metacognitive accuracy through deficit-oriented prompts (towards increased overestimation) based on confidence intervals as well as a small to moderate effect size, but significance testing for mean difference does not confirm these initial observations. We are hesitant to express a conclusion, but assume that if there was an effect of prompting on information processing and the formation of a metacognitive judgment, it might merely be a small effect. Other studies that addressed the effects of context-free and content-specific have found that context-free prompts are more effective than context-free prompts (e.g., Kramarski & Kohen, 2017). Having applied context-free prompts, our results are congruent with these studies.

In our study, we observed overall positive values in metacognitive accuracy in all groups implying an overestimation of judgment of comprehension. This finding is similar to a general and stable *overconfidence effect* as addressed by others (e.g., Gigerenzer, Hoffrage, & Kleinbölting, 1991; Koriat, Lichtenstein, & Fischhoff, 1980). This effect is said to occur when 'confidence judgments are larger than the relative frequencies of the correct answers' (Gigerenzer et al., 1991, p. 506). The *overconfidence effect* is generally observed when making a judgment after having answered a performance question. That is in contrast to our study design, in which a judgment about comprehension was made before answering performance questions.

The observed overestimation might also be a result of the difficult learning test that was designed to particularly obtain test results in and around the centre of the performance spectrum to avoid ceiling effects. A difference between judgment of comprehension and learning performance is not unexpected. Naturally, the absolute values of discrepancy in our study need to be viewed in a different light compared to classrooms assessments in which learners are supposedly more familiar with learning standards set by teachers. We would expect less overall discrepancy between judgment of comprehension and learning performance in settings in which learning goals and criteria for assessment are communicated transparently to learners (e.g., Bol et al., 2012).

The impact of metacognitive prompting on learning performance and self-evaluation

In this study, we compared the impact of resource-oriented and a deficit-oriented prompting on various measurements of learning. Contrary to previous findings (e.g., Berthold et al., 2007; Dori et al., 2018; Großschedl & Harms, 2013; Peters & Kitsantas, 2010), we did not observe improvements in learning performance. In accordance with the finding, that prompts need to be used regularly to be an effective tool to improve learning performance (Moser et al., 2017), we observe no immediate impact of prompts on learning performance. Beyond, we did not observe differences in satisfaction with learning, judgment of comprehension or the prediction of prolonged comprehension between both types of metacognitive questions.

Effectiveness of prompting and metacognitive activity in an online environment

In this study, the effectiveness of prompting was measured using a self-reported question regarding the use frequency of a question. Participants in the control group did not receive an introduction on metacognition and were likely to be unfamiliar with the term ‘metacognitive question’ in the self-report question. A short explanation on metacognitive question was integrated. We believe this short explanation led participants to report the use of a metacognitive question retrospectively and does not reflect metacognitive activity itself. In our view, it is likely that pre-service teachers in the control group did not apply the prompts deliberately and consciously as intended through prompting, and the measurement of use frequency is likely to be invalid. This is supported by the finding that the comparison of metacognitive accuracy with use frequency as covariate yielded no findings, suggesting that use frequency has no impact on metacognitive accuracy.

However, the effectiveness of prompting (how often and how extensive a metacognitive question is used) is likely to play a major role in its impact on metacognitive accuracy. We would like to raise the question whether learning environment impacts the effectiveness of prompts. Although prompts are widely investigated in online learning settings, they might have restricted effectiveness compared to classroom or group settings in which learning might be more standardised. For example, the use of metacognitive questions in a classroom might be instructed verbally, a certain time span might be specifically assigned to answer these questions and instructors could clarify task instruction in case of misunderstandings. Potential difference between learning settings could be a future study objective.

The role of extreme values in statistical analyses

We chose a conservative way of handling extreme values in this study (i.e., eliminating all extreme values that exceeded the threshold of two standard deviations from the mean value of each dependent variable). This was done to ensure that assumptions of the respective statistical test are met and to account for validity constraints that naturally accompany an online survey during the COVID19 pandemic with lock down restrictions. Indeed, there are good reasons to address extreme values. The extreme values that we observe might represent students that over- or underestimate their own performance particularly divergently from the average student (or perform particularly low or high) in the classroom. In naturalistic learning environments, these students might need individual support or feedback. In this study, we could not include these students due to our necessary statistical decisions and point out that those are the students that might benefit the most from prompts intended to improve metacognitive accuracy (see Kruger & Dunning, 1999). For future studies, we propose analyses of extreme values in naturalistic environments in an attempt to identify approaches to improve metacognitive accuracy in those groups who are particularly prone to over- and underestimation.

Limitations

As any study, this experimental study needs to be viewed in light of some limitations. Given that this study was carried out as an online survey, constraints in validity need to be addressed. Participants might have used additional aids to answer test questions despite being asked to refrain from doing so. Participants may have taken different amounts of time for reading and answering the question. Participants' motivation might have been diminished by the online survey in comparison to a classroom assessment. It needs to be questioned if they put appropriate effort into the task which we believe to be secured by the observed high interest in the topic. We addressed the anticipated limitation by choosing a conservative way of handling extreme values.

We did not identify in what depth students addressed and answered metacognitive questions. For instance, we were unable to observe whether pre-service took notes, how much time they spent answering the metacognitive questions and which specific contents they focused on. However, the depth and specificity in which the question is addressed probably contributes to the activation of memory traces and hence its impact on metacognitive accuracy and learning performance. A qualitative, laboratory research design with opportunities for participant observation or sufficient time for students' introspection might allow answering in what depth students addressed the metacognitive questions.

Conclusion and practical implications

Following the idea of the *cue-utilization approach* (Koriat, 1997), metacognitive accuracy is influenced by information processing during learning. With reference to the *embedded-processes model* (Cowan, 1988), we argued that attentional processes are guided through prompting. Through guided prompting, learners place their attention on specific information during the learning process. The information will be taken into account when forming a judgment of comprehension and hence impact metacognitive accuracy. Based on our analyses, we can neither confirm nor reject our hypothesis but assume small effects of prompting on metacognitive accuracy if any. Learning performance and judgment of comprehension were not impacted by the use of resource-oriented and deficit-oriented prompting. Other measurements of self-evaluation (i.e., satisfaction with learning outcome and prediction about prolonged comprehension) were not influenced by prompting either. Results are viewed in the background of online learning which might have restricted effectiveness of their implementation.

To increase the use of resource-oriented and deficit-oriented questions, we would like to address some practical considerations. These considerations are needed because addressing resources and deficits in an objective way might offer learners a range of opportunities for their academic development. Identifying gaps in comprehension is the key to filling those gaps, which in turn can lead to a manifestation of resources in the future. Finding a style of managing own resources and deficits and cultivating appropriate regulation of the attendant emotions can be seen as a goal of metacognition itself as much as it can be viewed as an opportunity for academic development. Recommendations on how to address deficits and mistakes specifically might include: a) making deficit-oriented prompts transparent and explaining how these questions are intended to improve metacognitive processes, b) communicate learning goals transparently, c) applying additional prompts to provide opportunities to overcome lacks in comprehension, d) acknowledging potential negative emotions that might be involved and e) creating an environment in which learners are encouraged to contribute openly and freely to classroom discussions and in which mistakes are not viewed as personal failures. Instead, a stance should be hold that supports the idea of deficits and mistakes being inherent to learning which offer the opportunity for development. We view the promotion of pleasure in understanding one's own thinking as the key to teaching metacognitive skills. Developing metacognitive skills and improving metacognitive accuracy is a long-term process to which resource-oriented and deficit-oriented prompts might contribute. For future research in this field we identify the need for long-term studies investigating efficiency of prompts on cognitive and emotional criterions.

Statements and Declarations

Supplementary Information

The online version contains supplementary material available at https:// doi.org/ 10. 1007/ s10212- 023- 00747-9.

Acknowledgements

We would like to thank Maren Flottmann, Britta Moers and Lara Tolentino for their professional proof-reading.

Authors' contribution

The authors agree with the content of this work and give their consent to its submission. The authors made the following contributions: SE: conceptualization & study design, acquisition, data analysis, data interpretation, original draft preparation and writing; JG: resources, revision and supervision.

Funding

Open Access funding enabled and organized by Projekt DEAL. This study was conducted using resources within the working group of Prof. Großschedl at the Institute for Biology Education, Faculty of Mathematics and Natural Sciences, University of Cologne, Cologne, Germany. This study received no specific funding.

Data availability

Data are openly available in https://osf.io/ykzvg/.

Ethics approval

The conduction of this online study involving human participants was in accordance with the ethical standards set in the declaration of Helsinki from 2013, the ethical guidelines of the German Psychological Society (DGPs) and the institutional ethical standards

Conflict of interest

The authors declare no competing interests.

References

Bannert, M., & Mengelkamp, C. (2013). Scaffolding hypermedia learning through metacognitive prompts *International handbook of Metacognition and Learning Technologies* (pp. 171-186): Springer. DOI: 10.1007/978-1-4419-5546-3_12

Bannert, M., & Reimann, P. (2012). Supporting self-regulated hypermedia learning through prompts. *Instructional Science, 40*(1), 193-211. DOI: 10.1007/s11251-011-9167-4

Bannert, M., Sonnenberg, C., Mengelkamp, C., & Pieger, E. (2015). Short-and long-term effects of students' self-directed metacognitive prompts on navigation behavior and learning performance. *Computers in Human Behavior, 52*, 293-306. DOI: 10.1016/j.chb.2015.05.038

Berthold, K., Nückles, M., & Renkl, A. (2007). Do learning protocols support learning strategies and outcomes? The role of cognitive and metacognitive prompts. *Learning and Instruction, 17*(5), 564-577. DOI: 10.1016/J.LEARNINSTRUC.2007.09.007

Bol, L., Hacker, D. J., Walck, C. C., & Nunnery, J. A. (2012). The effects of individual or group guidelines on the calibration accuracy and achievement of high school biology students. *Contemporary Educational Psychology, 37*(4), 280-287.

Cowan, N. (1988). Evolving conceptions of memory storage, selective attention, and their mutual constraints within the human information-processing system. *Psychological Bulletin, 104*(2), 163. DOI: 10.1037/0033-2909.104.2.163

Daumiller, M., & Dresel, M. (2019). Supporting self-regulated learning with digital media using motivational regulation and metacognitive prompts. *The Journal of Experimental Education, 87*(1), 161-176. DOI: 10.1080/00220973.2018.1448744

de Boer, H., Donker, A. S., Kostons, D. D., & van der Werf, G. P. (2018). Long-term effects of metacognitive strategy instruction on student academic performance: A meta-analysis. *Educational Research Review, 24*, 98-115. DOI: 10.1016/j.edurev.2018.03.002

Devolder, A., van Braak, J., & Tondeur, J. (2012). Supporting self-regulated learning in computer-based learning environments: systematic review of effects of scaffolding in the domain of science education. *Journal of Computer Assisted Learning, 28*(6), 557-573. DOI: 10.1111/j.1365-2729.2011.00476.x

Donker, A. S., De Boer, H., Kostons, D., Van Ewijk, C. D., & van der Werf, M. P. (2014). Effectiveness of learning strategy instruction on academic performance: A meta-analysis. *Educational Research Review, 11*, 1-26. DOI: 10.1016/j.edurev.2013.11.002.

Dori, Y. J., Avargil, S., Kohen, Z., & Saar, L. (2018). Context-based learning and metacognitive prompts for enhancing scientific text comprehension. *International Journal of Science Education, 40*(10), 1198-1220. DOI: 10.1080/09500693.2018.1470351

Döring, N., & Bortz, J. (2016). Forschungsmethoden und Evaluation. *Wiesbaden: Springerverlag*.

Dunlosky, J., & Lipko, A. R. (2007). Metacomprehension: A brief history and how to improve its accuracy. *Current Directions in Psychological Science, 16*(4), 228-232.

Dunning, D., Heath, C., & Suls, J. M. (2004). Flawed self-assessment: Implications for health, education, and the workplace. *Psychological Science in the Public Interest, 5*(3), 69-106. DOI: 10.1111/j.1529-1006.2004.00018.x

Engelmann, K., Bannert, M., & Melzner, N. (2021). Do self-created metacognitive prompts promote short-and long-term effects in computer-based learning environments? *Research and Practice in Technology Enhanced Learning, 16*(1), 1-21. DOI: 10.1186/s41039-021-00148-w

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American Psychologist, 34*(10), 906. DOI: 10.1037/0003-066X.34.10.906

Gardner, M. J., & Altman, D. G. (1986). Confidence intervals rather than P values: estimation rather than hypothesis testing. *Br Med J (Clin Res Ed), 292*(6522), 746-750.

Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models: a Brunswikian theory of confidence. *Psychological Review, 98*(4), 506. DOI: 10.1037/0033-295x.98.4.506

Großschedl, J., & Harms, U. (2013). Effekte metakognitiver Prompts auf den Wissenserwerb beim Concept Mapping und Notizen Erstellen. [Effects of metacognitive prompts on knowledge acquisition in concept mapping and note taking]. *Zeitschrift für Didaktik der Naturwissenschaften*, *19*, 375-395.

Haller, E. P., Child, D. A., & Walberg, H. J. (1988). Can comprehension be taught? A quantitative synthesis of “metacognitive” studies. *Educational Researcher, 17*(9), 5-8. DOI: 10.3102/0013189X017009005

Hemmerich, W. (2016). StatistikGuru: Rechner zur Adjustierung des α-Niveaus. Retrieved from https://statistikguru.de/rechner/adjustierung-des-alphaniveaus.html

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, 65-70.

Hübner, S., Nückles, M., & Renkl, A. (2006). *Prompting cognitive and metacognitive processing in writing-to-learn enhances learning outcomes.* Paper presented at the Proceedings of the Annual Meeting of the Cognitive Science Society.

Kim, N. J., Belland, B. R., & Walker, A. E. (2018). Effectiveness of computer-based scaffolding in the context of problem-based learning for STEM education: Bayesian meta-analysis. *Educational Psychology Review*, *30*(2), 397-429. DOI: 10.1007/S10648-017-9419-1

King, J. F., Zechmeister, E. B., & Shaughnessy, J. J. (1980). Judgments of knowing: The influence of retrieval practice. *The American journal of psychology,* 329-343.

Knippers, R. (2017). Epigenetik *Eine kurze Geschichte der Genetik* (pp. 327-354): Springer.

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*(2), 155-163. DOI: 10.1016/j.jcm.2016.02.012

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

Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human learning and memory, 6*(2), 107. DOI: 10.1037/0278-7393.6.2.107

Kramarski, B., & Kohen, Z. (2017). Promoting preservice teachers' dual self-regulation roles as learners and as teachers: Effects of generic vs. specific prompts. *Metacognition and Learning, 12*(2), 157-191. DOI: 10.1007/s11409-016-9164-8

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. DOI: 10.1037//0022-3514.77.6.1121

Leiner, D. J. (2022). SoSci Survey (Version 3.3.20) [Computer software]. Available at https://www.soscisurvey.de

Marsh, H. W., & Hattie, J. (1996). Theoretical perspectives on the structure of self-concept. In B. A. Bracken (Ed.), *Handbook of self-concept: Developmental, social, and clinical considerations* (pp. 38–90). John Wiley & Sons.

Mazzoni, G., & Cornoldi, C. (1993). Strategies in study time allocation: Why is study time sometimes not effective? *Journal of experimental psychology: General, 122*(1), 47.

Mazzoni, G., Cornoldi, C., & Marchitelli, G. (1990). Do memorability ratings affect study-time allocation? *Memory & Cognition, 18,* 196-204.

McCarthy, K. S., Likens, A. D., Johnson, A. M., Guerrero, T. A., & McNamara, D. S. (2018). Metacognitive overload!: Positive and negative effects of metacognitive prompts in an intelligent tutoring system. *International Journal of Artificial Intelligence in Education, 28*(3), 420-438. DOI: 10.1007/s40593-018-0164-5

Metcalfe, J., & Finn, B. (2008). Evidence that judgments of learning are causally related to study choice. *Psychonomic Bulletin & Review, 15*(1), 174-179. DOI: 10.3758/pbr.15.1.174

Metcalfe, J., & Kornell, N. (2005). A region of proximal learning model of study time allocation. *Journal of memory and language, 52*(4), 463-477.

Miller, T. M., & Geraci, L. (2014). Improving metacognitive accuracy: How failing to retrieve practice items reduces overconfidence. *Consciousness and cognition, 29*, 131-140.

Mitchum, A. L., Kelley, C. M., & Fox, M. C. (2016). When asking the question changes the ultimate answer: Metamemory judgments change memory. *Journal of experimental psychology: General, 145*(2), 200.

Moser, S., Zumbach, J., & Deibl, I. (2017). The effect of metacognitive training and prompting on learning success in simulation-based physics learning. *Science Education, 101*(6), 944-967. DOI: 10.1002/SCE.21295

Nelson, T. O., & Dunlosky, J. (1991). When people's judgments of learning (JOLs) are extremely accurate at predicting subsequent recall: The "delayed-JOL effect". *Psychological Science, 2*(4), 267-271.

Ocay, A. B. (2019). *Investigating the Dunning-Kruger effect among students within the contexts of a narrative-centered game-based learning environment.* Paper presented at the Proceedings of the 2019 2nd International Conference on Education Technology Management.

Quade, D. (1967). Rank analysis of covariance. *Journal of the American Statistical Association, 62*(320), 1187-1200.

Petermann, F., & Schmidt, M. H. (2006). Ressourcen-ein Grundbegriff der Entwicklungspsychologie und Entwicklungspsychopathologie? *Kindheit und Entwicklung, 15*(2), 118-127. DOI: 10.1026/0942-5403.15.2.118

Peters, E., & Kitsantas, A. (2010). The effect of nature of science metacognitive prompts on science students' content and nature of science knowledge, metacognition, and self-regulatory efficacy. *School Science and Mathematics, 110*(8), 382-396. DOI: 10.1111/j.1949-8594.2010.00050.x

Prinz-Weiß, A., Lukosiute, L., Meyer, M., & Riedel, J. (2022). The role of achievement emotions for text comprehension and metacomprehension. *Metacognition and Learning*, 1-27.

Rawson, K. A., Dunlosky, J., & Thiede, K. W. (2000). The rereading effect: Metacomprehension accuracy improves across reading trials. *Memory & Cognition, 28*, 1004-1010.

Rhodes, M. G., & Castel, A. D. (2009). Metacognitive illusions for auditory information: Effects on monitoring and control. *Psychonomic Bulletin & Review, 16*(3), 550-554. DOI: 10.3758/PBR.16.3.550

Rhodes, M. G., & Tauber, S. K. (2011). The influence of delaying judgments of learning on metacognitive accuracy: a meta-analytic review. *Psychological bulletin, 137*(1), 131. DOI: 10.1037/a0021705

Schraw, G. (1998). Promoting general metacognitive awareness. *Instructional Science, 26*(1), 113-125. DOI: 10.1023/A:1003044231033

Schraw, G., & Moshman, D. (1995). Metacognitive theories. *Educational Psychology Review, 7*(4), 351-371. DOI: 10.1007/BF02212307

Serra, M. J., & Dunlosky, J. (2010). Metacomprehension judgements reflect the belief that diagrams improve learning from text. *Memory, 18*(7), 698-711.

Shavelson, R. J., Hubner, J. J., & Stanton, G. C. (1976). Self-concept: Validation of construct interpretations. *Review of Educational Research, 46*(3), 407-441. DOI: 10.2307/1170010

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychological Science 22(11), 1359 - 1366. DOI: 10.1177/0956797611417632

Thiede, K. W., Anderson, M., & Therriault, D. (2003). Accuracy of metacognitive monitoring affects learning of texts. *Journal of educational psychology, 95*(1), 66.

Thiede, K. W., Dunlosky, J., Griffin, T. D., & Wiley, J. (2005). Understanding the delayed-keyword effect on metacomprehension accuracy. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*(6), 1267.

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.

Undorf, M., Söllner, A., & Bröder, A. (2018). Simultaneous utilization of multiple cues in judgments of learning. *Memory & Cognition, 46*, 507-519.

van Alten, D. C., Phielix, C., Janssen, J., & Kester, L. (2020). Effects of self-regulated learning prompts in a flipped history classroom. *Computers in Human Behavior, 108*, 106318. DOI: 10.1016/j.chb.2020.106318

Van den Boom, G., Paas, F., Van Merrienboer, J. J., & Van Gog, T. (2004). Reflection prompts and tutor feedback in a web-based learning environment: Effects on students' self-regulated learning competence. *Computers in Human Behavior, 20*(4), 551-567. DOI: 10.1016/j.chb.2003.10.001

Vangsness, L., & Young, M. E. (2021). More isn’t always better: when metacognitive prompts are misleading. *Metacognition and Learning, 16*(1), 135-156.

Veenman, M. V. (2012). Metacognition in Science Education: Definitions, Constituents, and Their Intricate Relation with Cognition *Metacognition in Science Education* (pp. 21-36): Springer. DOI: 10.1007/978-94-007-2132-6_2

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA statement on p-values: context, process, and purpose. *The American Statistician, 70*(2), 129-133.

Zheng, L. (2016). The effectiveness of self-regulated learning scaffolds on academic performance in computer-based learning environments: A meta-analysis. *Asia Pacific Education Review*, *17*(2), 187-202. DOI: 10.1007/s12564-016-9426-9

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. DOI: 10.1080/03057267.2013.847261

3.3 Study III: Promoting Self-Evaluation through Physical Exercise

Study III is not yet published.

> Elsner, S., Fränkel, S., Aschermann, E., & Großschedl, J. (unpublished). Strength- and Flexibility-Based Physical Exercise: An Experimental Study on Acute Effects on Attention, Self-Evaluation, and Emotional Responses in Children with ADHD.

3.3.1 Study III: Summary, Research Questions and Methodological Approach

Strength- and Flexibility-Based Physical Exercise: An Experimental Study on the Acute Effects on Attention, Self-Evaluation, and Emotional Responses in Children with ADHD. Study III examines whether the accuracy of self-evaluation in children with ADHD is affected by exercise-induced physiological changes. The "positive illusory bias" is a phenomenon observed in children with ADHD (Hoza et al., 2002). It refers to overly positive self-evaluations of competence in children with ADHD compared to their peers without ADHD (Hoza et al., 2004). These overly positive self-evaluations may result from cognitive impairments, i.e. limitations in working memory and attentional control, which are core symptoms of ADHD (e.g., McQuade et al., 2017). Information may not be kept in an "active" state that is required to form accurate self-evaluations (see, embedded processing model of working memory and cue-utilization framework, Cowan, 1999; Koriat, 1997). As a result, this information cannot be used to form self-evaluations, and self-evaluations become inaccurate. Cognitive functions have been shown to be positively impacted by physical exercise, even beyond other non-pharmacological interventions (Lambez et al., 2020). For instance, increased levels of noradrenaline and dopamine, as well as increased blood pressure and blood flow, are believed to alter brain functioning (e.g., Herold et al., 2019; Mulser & Moreau, 2023). The present study argues that attentional processes may be influenced by the physiological response to acute physical exercise, and that the alterations lead to enhanced accuracy in self-evaluation. Accordingly, this study pursues two objectives. First, it aims to examine potential effects of strength-based and flexibility-based physical exercise on attentional processes. Second, it investigated whether these effects influence self-evaluations.

This study addresses the following research questions:

> **RQ1:** Does acute physical exercise improve attentional performance in children with ADHD?
>
> **RQ2:** How do strength-based and flexibility-based physical exercises affect attentional performance?
>
> **RQ3:** If physical exercise alters attentional performance, does this improvement lead to more accurate self-evaluations?

To address these research questions, a within-subjects study with 24 children with ADHD was conducted. Each participant underwent strength-based training, flexibility-based training, and a control training session of approximately 30 minutes. Prior to the training and immediately after the training, participants completed an adapted version of the Eriksen Flanker Task (Eriksen & Eriksen, 1974; Ludyga et al., 2017). In this task, participants were shown fish pointing to the left or right and were asked to indicate, by keypress, the direction in which the middle fish was pointing. To determine whether physical exercise led to an increase in attention, results of the Eriksen Flanker Task in the three conditions were compared. To determine whether self-evaluation was affected by a potential increase in attention, two parameters were used. First, participants rated their task accuracy, i.e., “How often did you press the correct key?”, on a frequency scale including “never,” “seldom,” “sometimes,” “often,” and “always.” These responses were assigned a percentage value, e.g., the response “never” corresponded to “0%.” To calculate accuracy of self-evaluation, these values were subtracted from the objectively measured task accuracy, which was calculated as the proportion of correct responses out of all responses. Second, participants were asked to evaluate their reaction time: “How fast did you press the key?” with the response options “slower than before the training,” “equally fast as before the training,” “faster than before the training,” and “I don’t know.” To calculate the accuracy of self-evaluation, participants’ responses were categorised as “correctly rated,” “overestimated,” “underestimated,” and “no rating”. Self-evaluation results were analysed using descriptive statistics.

3.3.2 Study III: Own Contribution

The author of this dissertation designed and planned this study. The author also developed the training program with support from Armin Oster, and conducted participant recruitment with the assistance of Selina Faist, a student assistant who completed her thesis within the project. The study was also conducted by the author of this dissertation who was responsible for data preparation and analysis, as well as for writing the manuscript. The contributions of all authors can be found in the unpublished manuscript.

3.3.3 Study III: Unpublished Manuscript

Strength- and Flexibility-Based Physical Exercise: An Experimental Study on Acute Effects on Attention, Self-Evaluation, and Emotional Responses in Children with ADHD

Elsner, S., Fränkel, S., Aschermann, E., & Großschedl, J.

Abstract

Attention deficit hyperactivity disorder (ADHD) is a neurodevelopmental disorder, characterised by symptoms of inattentiveness, impulsivity, and hyperactivity. The effects of physical exercise on cognition, metacognition, and emotional responses are attracting increasing interest in ADHD research, with promising findings emerging for ADHD treatment. This study examined the impact of strength-based and flexibility-based exercises on attention, self-evaluation, and emotional responses. We examined the acute effects of 30-minute training sessions - strength vs. flexibility vs. control - in 24 children with ADHD. The results revealed no evidence for an effect of strength-based or flexibility-based training on attention, self-evaluation, or emotional responses. Discrepancies in ratings of emotional states were observed between children's and parents' perceptions. Compared to previous studies on endurance-based training, the types of exercise investigated here may not produce similar physiological responses or effects on attention. Training intensity may play a key role in inducing effects on attention.

Keywords: cognition, metacognition, training, parent, academic

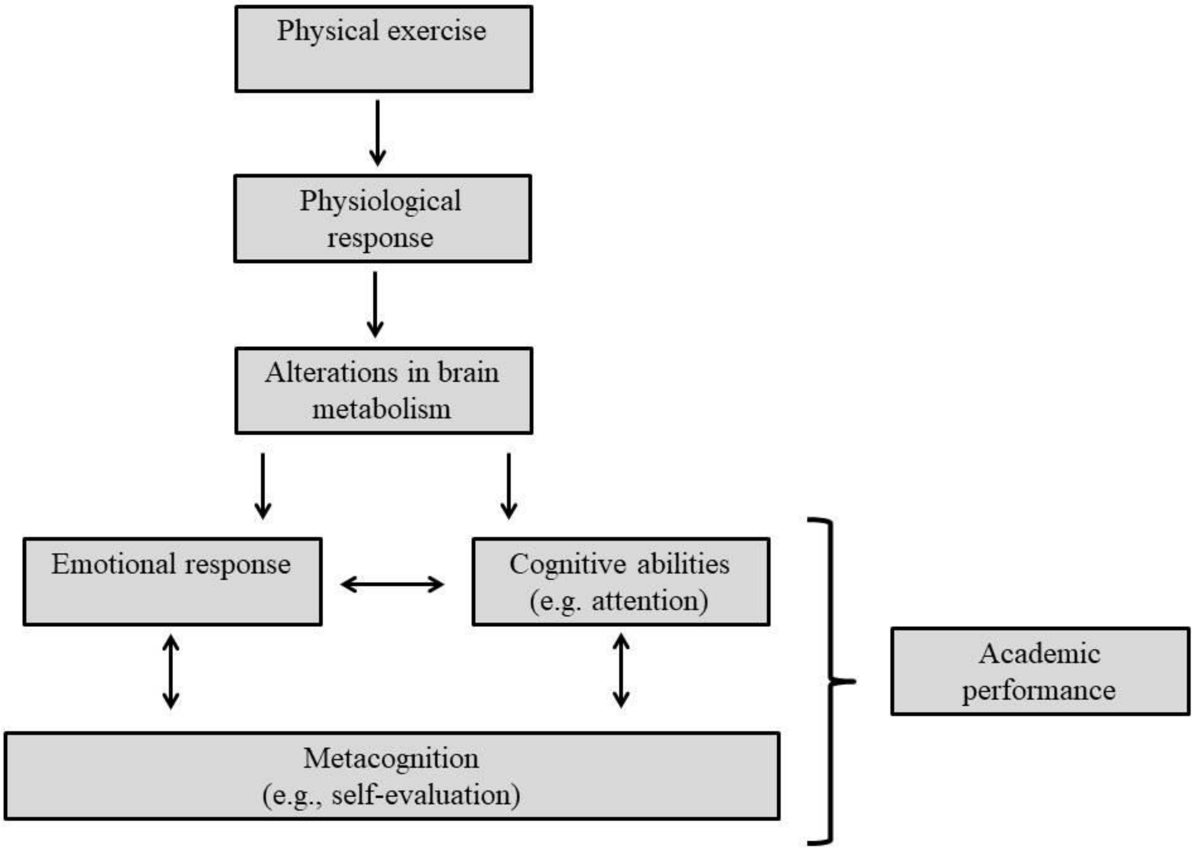
Background

ADHD in School Settings

Attention deficit hyperactivity disorder (ADHD) is a neurodevelopmental disorder affecting self-regulation. It is characterised by a behavioural pattern of inattentiveness, impulsivity, and hyperactivity (Faraone et al., 2021). Diagnostic criteria include "the presence of developmentally inappropriate levels of hyperactive-impulsive and/or inattentive symptoms for at least 6 months" with impairments in quality of life (Faraone et al., 2021, p. 793). Reduced cognitive functioning and a high level of psychological strain are, by definition, associated with ADHD (Faraone et al., 2021). The prevalence rate of ADHD lies between 6.1 to 9.4 % (Salari et al., 2023), indicating that an average of one to three children in a classroom of 30 students are diagnosed with ADHD. This high prevalence rate suggests that teachers are likely to encounter effects of ADHD symptoms in their classroom. Indeed, typical symptoms of ADHD manifest as barriers to effective learning. Academic underachievement is commonly observed (Arnold et al., 2020; Daley & Birchwood, 2010; Frazier et al., 2007). For example, teachers report that students with ADHD are less likely to reach their full potential (Kent et al., 2011). Children with ADHD perform below their predicted skill levels in reading, writing, and mathematics (Barry et al., 2002). In particular, reading tests reveal negative associations with ADHD symptoms (Frazier, 2007). Children with ADHD frequently struggle to meet scholastic expectations, such as submitting homework assignments on time (Langberg et al., 2016). Moreover, ADHD-related impairments can persist over time (Massetti et al., 2008). Teachers need to respond professionally to children's needs, enabling them to reach their potential despite – or even because of – their ADHD symptoms (Sedgwick et al., 2019). Interestingly, students' underachievement remains evident even when scholastic outcomes are controlled for intelligence (Daley & Birchwood, 2010; Loe & Feldman, 2007), suggesting that cognitive factors other than intelligence are more likely to explain academic underachievement. Specifically, impairments in the self-regulation of executive functions and motivation are believed to lead to decreased performance, characterised as diminished quality and quantity of task engagement (Sonuga-Barke, 2002, 2003). Specifically, inhibitory dysfunctions that drive cognitive and behavioural dysregulation (i.e., inattention) and a generalised delay aversion presumably contribute to this lack of task engagement (e.g., Sonuga-Barke, 2003). Task engagement, however, is crucial in school, and attentiveness is often required due to the formal nature of the learning setting. At the same time, rewards such as positive feedback are not always provided immediately. The dual pathway model suggests that an imbalance in dopamine-related pathways in brain metabolism is one of the causes of self-regulatory difficulties (Sonuga-Barke, 2002, 2003). Schools and researchers are interested in multimodal approaches supporting learners with ADHD in classrooms and addressing these imbalances (DuPaul et al., 2011). In addition to skill acquisition and instructional changes (e.g., DuPaul et al., 2011), the implementation of physical exercise programs may be a promising opportunity, as physical exercise can directly impact brain physiology (see Figure 1[7]).

Figure 1

Proposed Effect Mechanism of Physical Exercise on Academic Performance


Physical exercise
Physiological response
Alterations in brain metabolism
Emotional response
Cognitive abilities (e.g. attention)
Metacognition (e.g., self-evaluation)
Academic performance

[7] A similar line of reasoning can be found in Tomporowski et al. (2015).

Exercise-Induced Effects on Cognition

The acute effects of exercise are gaining interest in ADHD research (Christiansen et al., 2019; Grassmann et al., 2017). These studies contribute to a better understanding of already overwhelming evidence of positive long-term effects on cognitive functioning in general (Audiffren & André, 2019; Chaddock et al., 2012; Hillman et al., 2011; Kramer et al., 2006; Liang et al., 2022; Tomporowski, 2003b) and in children with ADHD (Lambez et al., 2020; Meßler et al., 2018). A myriad of biological components and physiological processes – including the dopamine system – likely underlie cognitive enhancement through physical exercise. Physical exercise has been shown to lead to the release of neuromodulatory molecules into the bloodstream. Changes in noradrenaline, adrenaline, cortisol, lactate, and dopamine levels after physical activity have been observed in adults (Herold et al., 2019; Van Hall et al., 2009). Elevated levels of lactate, somatropin, insulin-like growth factors (IGF-2), testosterone, cortisol, and adrenaline were observed in children (Armstrong & Van Mechelen, 2017). These molecules can cross the blood-brain barrier and likely reach different brain regions. Here, neuromodulatory molecules may alter cognitive functions (e.g., Tomporowski et al., 2015). Additionally, increased blood pressure and blood flow likely foster rapid transport of neurochemicals and increase their availability in the brain (Mulser & Moreau, 2023; Skriver et al., 2014). The current state of scientific evidence is still in need of clarity regarding molecular and physiological mechanisms, which can only be partly addressed in this study. However, there is already a strong link between physical exercise and attentional processes (e.g., Basso & Suzuki, 2017; Tomporowski, 2003a), scholastic performance in general (Castelli et al., 2007; Caterino & Polak, 1999; De Greeff et al., 2018; Hillman et al., 2011; Moreau et al., 2017; Singh et al., 2019), and symptom reduction in children with ADHD (Christiansen et al., 2019; Den Heijer et al., 2017; Neudecker et al., 2019). A recent meta-analysis found that physical exercise yielded the largest effects when compared to other non-pharmaceutical interventions in children with ADHD (Lambez et al., 2020). Considering concerns about medication, such as the rate of non-responders, parents' potential reservations, its low impact on cognitive functions and school performance (Coghill et al., 2014; Daley et al., 2014; Jangmo et al., 2019; Kortekaas-Rijlaarsdam et al., 2019), physical exercise is a promising approach in ADHD treatment.

Exercise-Type and Effects on Cognition

We assume that the neurochemical release after acute exercise, and hence its effects on cognition, depends on the type of exercise. Exercise-induced effects in children with ADHD have mainly been investigated in response to endurance-based training (e.g., Neudecker et al., 2019). Endurance-based training aims to increase aerobic fitness (Armstrong & Van Mechelen, 2017). Indeed, aerobic exercise seems to impact cognitive functions positively. After a 30-minute run or cycling session, performance in a selective attention task improved (Chang et al., 2012; Piepmeier et al., 2015). Previous studies have shown improvements in reaction times and task accuracy following endurance-based training (Ludyga et al., 2017; Ludyga et al., 2020; Medina et al., 2010; Pontifex et al., 2013). While the cognitive effects of endurance-based training have been fairly well studied, other forms of exercise – such as strength-based or flexibility-based training – have received less attention in ADHD research. However, a recent meta-analysis reported preliminary positive effects of such exercises on academic outcomes in non-ADHD populations (Robinson et al., 2023).

Strength-based training aims to increase muscular strength, while flexibility-based exercise aims to increase the range of motion (Armstrong & Van Mechelen, 2017). Both exercise types can serve as alternative options to movement patterns in endurance-based training like cycling and running, which potentially may be perceived as monotonous by children with ADHD. They offer a greater variety of movement patterns and likely provide more stimulating challenges, accommodating children’s sense of boredom, which is often anecdotally reported by care givers and children themselves. At the same time, both exercise types already represent natural movement patterns for children (e.g., in climbing). Strength-based and flexibility-based exercises likely cause different physiological responses. We expect that strength-based training provides a higher training stimulus, elicits greater physiological responses, and increases the release of neuromodulatory factors. Average heart rate and maximum heart rate are likely to increase after strength-based training resulting in greater blood flow and blood pressure compared to flexibility-based training. Strength-based training maybe perceived as more intense and effortful. Due to the hypothesised stronger physiological response to strength-based raining, we assume greater effects on cognition compared to flexibility-based training.

Physical Exercise and Self-evaluation in ADHD

The term “positive illusory bias” (PIB) describes differences between subjective ratings of competence given by children with ADHD and those provided by their parents or teachers (Hoza et al., 2004; Volz-Sidiropoulou et al., 2016; Owens et al., 2007). The PIB suggests that children with ADHD overestimate their competencies compared to typically developing children (Hoza et al., 2004), particularly in areas where they experience impairments (Hoza et al., 2002).The PIB has also been observed in relation to daily life activities and social competencies (Emeh et al., 2018; McQuade et al., 2017). Importantly, a critical review points out that the PIB is not necessarily specific to ADHD but may be a “function of general impairment – ADHD-related or otherwise” (Owens et al., 2007). Either way, it is necessary to address overly positive self-evaluations in children with ADHD, as accurate self-evaluations drive appropriate learning behaviour (e.g., Metcalfe & Finn, 2008).

The accuracy in self-evaluation may be impacted by physical exercise, but, to our knowledge, this has not yet been investigated. Cognitive impairments likely contribute to inaccurate self-evaluations (for alternative explanations, see McQuade et al., 2017). If attention cannot be sustained during task execution, it seems logical that task-relevant information will be processed less coherently. However, the processing of task-relevant information is necessary to accurately self-evaluate (e.g., see cue utilization framework; Koriat, 1997). If physical exercise can induce improvements in cognitive processing and attention, more predictive indicators of task-performance might also be processed and integrated into the formation of self-evaluation. Consequently, self-evaluations are likely to become more accurate. A similar line of reasoning can be in a narrative review on exercise-induced effects, primarily in typical developing children (Tomporowski et al., 2015). In this review, physical exercise is proposed to impact academic achievement through improvements in cognitive functions and metacognition. In our study, we test the hypothesis that self-evaluations become more accurate in response to physical exercise, alongside improvements in attentional parameters. We hypothesise that strength-based training has a greater effect on the accuracy of self-evaluation compared to flexibility-based training.

Physical Exercise and Emotional Response in ADHD

Studies investigating the effects of physical exercise in children with ADHD mainly focus on cognitive parameters and behavioural outcome measures (Den Heijer et al., 2017; Ng et al., 2017; Vysniauske et al., 2020). Effects on emotional response after physical exercise are less well investigated, although ADHD is strongly associated with emotional dysregulation

(Bunford et al., 2015; Christiansen et al., 2019; Graziano & Garcia, 2016). Some evidence suggests positive effects of acute exercise on emotional well-being in children and adults (Bigelow et al., 2021; Fritz & O'Connor, 2016). To our knowledge, there has not yet been an investigation of strength-based and flexibility-based training on emotional response in children with ADHD. We aim to investigate whether children's emotional states alter after physical exercise. We expect emotional responses to be influenced by the physiological response after physical exercise through neurotransmitter release. We expect feelings of pleasure and arousal to increase in both types of training, with strength-based training being more effective than flexibility-based training.

Study Aims

Our study aims to directly compare the impact of strength-based and flexibility-based training on attention and self-evaluation in children with ADHD. A secondary aim is to examine children's emotional responses and parental perceptions. We assume that the release of neuromodulatory substrates following acute exercise affects brain metabolism, resulting in improved attention and more accurate self-evaluations. We assume that strength-based training will have lager effects than flexibility-based on all depended variables due to increases physiological responses.

Material and Methods

Participants

Participants were recruited through flyers at local medical offices for child and youth psychotherapy and occupational therapy, in public places, at sports centers, and through private contacts. Participating children had to have been diagnosed with ADHD before their participation by a medical clinician or a psychological psychotherapist. The formal diagnosis was verified before participation in the study. An exclusion criterion was a diagnosis of autism spectrum disorder. Legal guardians were advised to consult their pediatrician regarding possible pre-existing conditions that might contraindicate participation in the study. A total of 26 participants took part in this experimental study. Two participants had to be excluded because they did not follow the exercise instructions closely enough to evoke a physiological response. The remaining participants were between 7 and 12 years old (M = 9.5 years, SD = 1.4). Nineteen children were male, and five were female. Fifteen children had a standard weight, six were slightly overweight, and one was overweight based on the age-adjusted body mass index (online calculator; BKK Gesundheit). Eleven children were taking regular medication to treat ADHD-related symptoms, while 13 children were not taking any medication. Children who took medication regularly either refrained from taking it on the testing days, or testing was scheduled to minimize the effects of medication. For this reason, testing was scheduled in the afternoon when the effects of the medication had worn off. Eighteen children were diagnosed with ADHD, and six were diagnosed with ADD. All participants and their legal guardians provided written informed consent before testing. Parents and children were free to withdraw from the study at any point of assessment. Participants received a reimbursement of €45 for their participation. All study procedures followed the guidelines of the Declaration of Helsinki (Association, 2001). This study was approved by the Ethics Committee of the Faculty of Human Sciences at the University of Cologne (identification number: SEHF0164).

Procedures

This experimental, within-subjects study comprised three approximately one-hour sessions, with a one-week interval between sessions. In cases of scheduling conflicts or unforeseen health issues, the interval was extended to two weeks (*n* = 3). Each participant took part in a strength-based session, a flexibility-based session, and a control session. The order of these three sessions was randomized. The first session began with welcoming the participants and providing them with an opportunity to ask questions about the study. The study procedures were explained, and legal guardians were asked to wait near the experimental room until the end of the session. During the first session, legal guardians completed a questionnaire to collect demographic information about the participating child. The study began with a practice version of the modified Eriksen Flanker Task, which measures inhibitory control (Eriksen & Eriksen, 1974; Ludyga et al., 2017). Subsequently, participants performed a strength-based, a flexibility-based, or an inactive control training for approximately 30 minutes. The physically inactive control training involved watching an age-appropriate documentary about animals. Following the strength-based and flexibility-based training, participants were asked to rate their perceived exertion using the Effort Scale Sport (German: ‘Anstrengungsskala Sport’) on a scale from 1 (‘not at all exhausting’) to 10 (‘so exhausting, that I need to stop’; Buesch et al., 2021). Before and after all training sessions, participants’ affective states were measured. After training, participants performed a modified Eriksen Flanker Task for approximately 12 minutes and then self-evaluated their performance. Participants wore an optical heart rate sensor (Polar Verity Sense) throughout the entire experiment. The entire study was conducted under the supervision of one researcher.

Training Characteristics

Participants exercised by watching and following videos of a strength-based and a flexibility-based training routine. Both videos were pre-recorded to ensure standardization of training instructions (see the attachment for the training routine). Visual and verbal cues were provided to guide the execution of exercises. The training sessions consisted of a 5-minute warm-up, a 20-minute main training session, and a 5-minute cool-down. The warm-up and cool-down were identical. The strength-based training included exercises such as squats, single-leg stands, and press-ups. The flexibility-based training included exercises such as the cat-cow movement in an all-fours position, the sprinter's stretch with a forward reach, and forward bends. Both training sessions included exercises adapted from the German Motor Test for ages 6 to 18 years (Boes, 2017; Deutscher Motorik-Test). The inclusion of these exercises aimed to gather diagnostic information about strength and flexibility levels, i.e., standing long jumps, left-right skips, press-ups, and forward bends. Both training sessions were designed to be age-appropriate for children. The warm-up was designed to prepare muscle groups and joints that are particularly prone to injuries in children, such as upper body, including shoulders and wrists (Faigenbaum et al., 2009). The exercises did not include additional weights but consisted solely of bodyweight activities. Compared to adult workouts, the load duration was shortened, the number of repetitions was reduced, and transitions between exercises were quicker. Figurative language was used to instruct, e.g., 'Superman breathing,' 'knee hug,' 'robot,' and 'zombie.' Children could choose to repeat their favorite exercise at the end of the training sessions.

Modified Eriksen Flanker Task

The Eriksen Flanker Task is a measure of information processing and inhibitory control (Eriksen & Eriksen, 1974). In this study, a modified version of the Eriksen Flanker Task was implemented. In a computerised task, participants had to respond to a centrally displayed fish (stimulus) within a horizontal row of five fish by pressing a key (see 3). Participants were instructed to press the "M" key on a keyboard (British layout) if the central fish pointed to the right. They were instructed to press the "Z" key if the central fish pointed to the left. The trials were either congruent or incongruent. In congruent trials, the centrally displayed fish was flanked by fish pointing in the same direction. In incongruent trials, the centrally displayed fish was flanked by fish pointing in the opposite direction. Each trial began with a fixation cross in the centre of the screen for 1000ms, followed by the stimulus (congruent/incongruent and right/left) for 1000ms, and a feedback display for 1500ms. The feedback display showed the accuracy of the response to the stimulus (correct or incorrect) and the reaction time for correct responses in milliseconds. The trials were presented in randomized order. In accordance with Ludyga et al. (2017), four blocks of 40 trials each were presented. Each block consisted of ten trials per trial type, i.e., congruent-left, congruent-right, incongruent-left, incongruent-right. The task was preceded by 12 practice trials. A thorough practice phase, consisting of two blocks of 20 trials each, was conducted prior to the training sessions to ensure familiarity with the task. Instructions were given verbally by the instructor and shown on the screen. This modified Eriksen Flanker Task was created with E-Prime 3.0 Software (Psychology Software Tools, Pittsburgh, PA).

Self-Evaluation

Self-evaluation was assessed using two instruments. First, we assessed perceived accuracy of task accuracy. To do this, participants were asked to rate how often they thought they had pressed the correct key. Participants responded using a thumb-based scale frequency labels (‘never,’ ‘seldom,’ ‘sometimes,’ ‘often,’ and ‘always’). Second, we assessed perceived accuracy of reaction time. To do this, participants rated how quickly they thought they had pressed the key. Response options included ‘slower than before the training,’ ‘as fast as before the training,’ ‘faster than before the training,’ and ‘I don’t know.’

Emotional response

Before and after all training sessions, participants’ affective states in the dimensions of ‘pleasure,’ ‘arousal,’ and ‘dominance’ were assessed using Self-Assessment Manikins (Bradley & Lang, 1994). Parents’ ratings of their children’s emotional responses were measured. For this, we adapted the Self-Assessment Manikin scale to a visual analogue scale. Parents rated their children emotional response of ‘pleasure,’ ‘arousal’, and ‘dominance’ on a scale from 1 to 9, with increments of one point, for the remainder of the day. The ‘dominance’ dimension was excluded from analysis due to validity concerns based on feedback from parents and children. Participants were also asked to rate their level of enjoyment during the training sessions. For this, they rated their agreement with the statement: “This training/watching the video was fun” by selecting a thumb on a pictorial scale (1 = low agreement, 5 = high agreement).

Data Preparation and Data Analyses

Heart Rate

Heart rate was measured throughout the entire experiment. Mean heart rate values were calculated for each participant: [1] during the execution of the Erikson Flanker Task before training, [2] during training (excluding warm-up and cool-down), and [3] during the execution of the Eriksen Flanker Task after training. We also analysed the maximum heart rate, i.e., the highest value during training.

Eriksen Flanker Task

Task accuracy and reaction times were calculated for the Eriksen Flanker Task [1] prior to training and [2] after training. Only trials with reaction times greater than 200ms were analysed to account for anticipatory presses before children saw the stimulus. Task accuracy and reaction times were calculated for congruent (all fish pointed in the same direction) and incongruent (flanking fish pointed in the opposite direction) trials. Task accuracy was calculated as the proportion of correct responses out of all responses. Reaction time was calculated as the mean value. After training, the mean value was calculated for all trials in blocks 1 to 4, excluding practice trials. To account for variability in reaction times, we also calculated the standard deviation for each participant in the Erikson Flanker Task after training. We expected a speed-accuracy trade-off, meaning that reaction time might be increased in favour of more accurate responses and accuracy might be decreased in favour of faster responses because cognitive performance is reflected in both task accuracy and reaction time. No systematic response pattern was expected. To account for this speed-accuracy trade-off, we calculated the inverse efficiency score (IES), which allows the combined assessment of reaction time and accuracy (e.g. Yeung et al., 2020). For this, we divided reaction time by the proportion of correct responses, following Yeung et al. (2020). We also expected a typical congruency effect, meaning that children’s performance in incongruent trials is likely to be decreased compared to their performance in congruent trials (e.g., Eriksen & Eriksen, 1974).We accounted for this by subtracting IES values of incongruent trials from IES values of congruent trials. Task accuracy, reaction times, and standard deviations were compared between the three training conditions using repeated-measures statistical analyses.

Accuracy of Self-Evaluation

Accuracy of Self-evaluation of Task Accuracy

We aimed to assess the accuracy of self-evaluation regarding task accuracy. The accuracy of self-evaluation refers to the congruence of subjective assessment between the objective measurements. Task accuracy is defined as the proportion of correct responses relative to the total number of responses. Self-evaluation of task accuracy was measured with a thumb-based rating scale. Response options were matched to percentage values (see Table 3). We then subtracted objectively measured accuracy from the matched values to estimate the accuracy of self-evaluation. For example, if a participant rated their task accuracy as ’seldom’ but achieved an objective task accuracy of 50%, we calculated 25 % minus 50%. The differences between self-evaluations and objective task accuracy were analysed across the training conditions with inferential statistics.

Table 3

Assignment of Self-Evaluation Ratings to a Percentage Value

	never		seldom		Some-times		often		always
Responses	1	1,5	2	2,5	3	3,5	4	4,5	5
Percentage	0%	12,5%	25%	37,5%	50%	62,5%	75%	87,5%	100%

Accuracy of Self-evaluation of Reaction Time

Our aim was to assess the accuracy of self-evaluations of reaction time. Self-evaluation of reaction time was measured using four response options to the question: “How fast do you think you pressed the key?” The response options were: ‘slower than before the training,’ ‘as fast as before the training,’ ‘faster than before the training,’ and ‘I don’t know.’ First, we calculated the difference in reaction time between post-training and pre-training values. For each condition, a mean and standard deviation of reaction time was calculated across all participants (see Table 4). If a participant’s reaction time was more than one standard deviation below the mean difference, it was classified as ‘faster than before the training.’ If a participant’s reaction time was more than one standard deviation above the mean, it was classified as ‘slower than before the training’. A reaction time within one standard deviation of the mean was classified as ‘as fast as before the training.’ Subsequently, participants’ self-evaluations of reaction time were compared with the corresponding classification. Correct self-evaluations were defined as responses in which participants’ self-evaluations matched our classification. Overestimation was defined as responses in which participants rated their reaction time in a higher category. Underestimation was defined as responses in which children rated their reaction time in a lower category. The results were analysed using descriptive statistics.

Table 4

Mean and Standard deviation of Differences in Reaction Times across Training Conditions

	Mean in ms	Standard deviation in ms	Range within one standard deviation in ms
Strength-based training	5.29	48.91	-43.62 to 54.20
Flexibility-based training	-2.74	66.97	-69.71 to 64.23
Control training	13.52	43.38	-29.86 to 56.90

Statistical Analyses

Prior to applying inferential statistics, we checked whether the dependent variables met test assumptions. In cases of violations of the test assumptions, we applied non-parametric tests. To determine differences between training conditions (i.e., strength-based training = StrT, flexibility-based training = FlexT, and control training = ConT), we applied one-way ANOVAs for repeated measures or Friedman tests. Post-hoc tests are mentioned where applicable. Statistical analyses were carried out using IBM SPSS Statistics Version 29.0 (IBM Corp., 2021). The α-level was set to .05. Graphs were created with JASP . Effect sizes were either provided by SPSS or calculated using the website ‘psychometrica’ (Lenhard & Lenhard, 2022).

Results

Participants' Fitness Level

Children's fitness levels were assessed using strength and flexibility indicators of the German Motoric Test (Boes, 2017). Strength was measured by the distance of long jumps and the number of press-ups. The results were compared to a normative sample (Boes, 2017). Ten out of 24 children (41.7%) achieved age-typical results, i.e., values within one standard deviation below or above the mean. Seven children (29.2%) achieved results above one standard deviation in one discipline (i.e. long jump, press-ups). Seven children (29.2%) achieved results at least one standard deviation below in one of the two disciplines. Two of these children (8.3%) achieved results below two standard deviations in both disciplines. Flexibility was measured using forward bends. Twenty-two out of 24 children (91.7%) reached age-typical values. Two children each achieved values below and above two standard deviations (8.3% each).

Manipulation Check and Training Evaluation

To ensure that training led to an increase in physical activity as intended by the experimental design, we analysed differences in heart rate between StrT, FlexT, and ConT. We compared heart rate before, during, and after training. Table 5 shows descriptive data and inferential statistics. We did not find evidence for differences in mean heart rate at baseline between the three conditions: χ^2 (2) = 2.44, *p* = .296, Kendall's *W* = .053, via a Friedman Test. During training, mean heart rate differed between groups: χ^2 (2) = 42.35, *p* < .001, Kendall's *W* = .921. Post-hoc Wilcoxon signed-rank tests revealed higher mean heart rates during StrT compared to FlexT (*Z* = -4.20, *p* < .001, η^2 = 0.767), during StrT compared to the ConT (*Z* = -4.20, *p* <.001, η^2 = .767), and during FlexT compared to the ConT (*Z* = -3.89, *p* < .001, η^2 = 0.631). This increase in mean heart rate remained evident after training (χ^2 (2) = 9.25, *p* = .010, Kendall's *W* = .201), but only between StrT and FlexT (*Z* = -3.16, *p* = .002, η^2 = 0.434), and between StrT and the ConT (*Z* = -2.14, *p* = .032, η^2 = 0.199) and not between FlexT and the ConT (*Z* = -0.39, *p* = .700, η^2 = 0.006) as indicated by post-hoc Wilcoxon signed-rank tests. Maximum heart rate differed between conditions as calculated by a repeated measures ANOVA: *F* (2, 44) = 171.9, *p* < .001, $\eta_p{}^2$ = .887. Bonferroni post-hoc tests revealed higher maximum heart rate in StrT compared to FlexT (*p* < .001, *d* = 1.83), higher maximum heart rate in StrT compared to ConT (*p* < .001, *d* = 3.05), and in FlexT compared to the ConT (*p* < .001, *d* = 2.22). Ratings of applied effort were examined for differences between StrT and FlexT. A paired-samples *t*-test was used. Ratings of applied effort during training were higher in StrT than FlexT: t (23) = 2.94, p = .007, d = 0.62. Results indicated moderate levels of effort in both training types with mean values of 4.73 (*SD*: 2.42) in StrT and 3.35 (*SD*: 2.22) in FlexT (0 = low effort, 10 = high effort). Feelings of enjoyment were tested for differences between StrT, FlexT, and ConT. Results of the Friedman test revealed differences between conditions: χ^2 (2) = 7.86, p = .020, Kendall's W = .164. Post-hoc Wilcoxon signed-rank tests indicated less enjoyment during FlexT compared to ConT; Z = -2.52, p = .012; η^2 = 0.265. Means and standard deviations indicated high levels of enjoyment overall: StrT: *M*: 4.17, *SD*: 1.26; FlexT: *M*: 3.90, *SD*: 1.42, ConT: *M*: 4.69, *SD*: 0.55 (1 = low enjoyment, 5 = high enjoyment).

Table 5

Heart Rate Before, During, and After Training

		N	Mean	SD	Inferential statistics
HR before training	StrT	23	87.3	10.2	χ^2 (2) = 2.44
	FlexT	24	88.4	9.6	p = .296
	ConT	24	88.7	11.3	Kendall's W = .053
HR during training	StrT	23	125.6	11.3	χ^2 (2) = 42.35
	FlexT	24	101.3	6.3	p < .001
	ConT	24	87,3	10.6	Kendall's W = .921
HR after training	StrT	23	94.8	9.25	χ^2 (2) = 9.25
	FlexT	24	88.4	7.2	p = .010
	ConT	24	90.2	11.4	Kendall's W = .201
Maximum HR during training	StrT	23	161.5	12.9	F (2, 44) = 171.9
	FlexT	24	131.9	11.6	p < .001
	ConT	24	110.6	10.1	η_p^2 = .887

Note. StrT = strength-based training, FlexT = flexibility-based training, ConT = control training, HR = heart rate. Inferential statistics show the results of overall comparisons between the three conditions. Detailed post-hoc analyses can be found above.

Effects of Exercise on Attention

We tested whether strength-based and flexibility-based training affected attention in a subsequent Erikson Flanker Task. We analysed reaction time, task accuracy, and the inverse efficiency score (IES), which combines reaction time and task accuracy. Table 6 shows descriptive data on attentional parameters. We did not find differences in the IES between strength-based, flexibility based, and the control training in congruent through Friedman Tests (χ^2 (2) = 1.41, *p* = .494; Kendall's *W* =.029) and incongruent trials (χ^2 (2) = 3.85, *p* = .146; Kendall's *W* = .080). Nor did we observe differences between the three training conditions when accounting for the congruency effect by subtracting the IES of congruent trials from IES of the incongruent trials though a repeated measures ANOVA (*F* (2, 46) = 0.28; *p* = .754, η_p^2 = .012). Reaction time and task accuracy were analysed separately. We did not observe differences in reaction time between strength-based, flexibility-based, and the control training in congruent trials (*F* (2, 46) = 0.03, *p* = .970, η_p^2 = .001) or in incongruent trials (*F* (2, 46) = 0.24, *p* = .786, η_p^2 = 0.01) through repeated measures ANOVA. We did not observe differences in task accuracy between strength-based, flexibility-based, and the control training in congruent trials (*F* (2, 46) =.046; *p* = .956; η_p^2 = 0.002), nor in incongruent trials (*F* (2, 26) = 0.057, *p* = .945, η_p^2 = 0.002) trough repeated measures ANOVA. Figures 1 and 2 show the task accuracy in all three conditions. We tested whether the variability of reaction time differed after strength-based training, flexibility-based training, and control training. The standard deviation was not found to differ in congruent (*F* (2, 46) = 0.73; *p* = .488) and incongruent trials (*F* (2, 46) = 0.39; *p* = .680), as indicated by a repeated measures ANOVAs. We investigated whether strength-based, flexibility-based, and control training differed in their impact on task effort and task enjoyment of the Erikson Flanker Task. We did not observe a difference in the effort put into the task between training conditions in a Friedman Test (χ^2 (2) = 2.91, *p* = .233; Kendall's *W* = .061), with medians of 5 in the strength-based and flexibility-based training, and a median of 4 in the control training on a scale from 1 to 5. We did not observe a difference in task enjoyment between the training conditions in a Friedman Test (χ^2 (2) = 2.75, *p* = .252; Kendall's *W* = .057), with medians of 4 after strength-based and control training, and a median of 5 after the flexibility-based training.

Figure 1

Task Accuracy in Congruent Trials

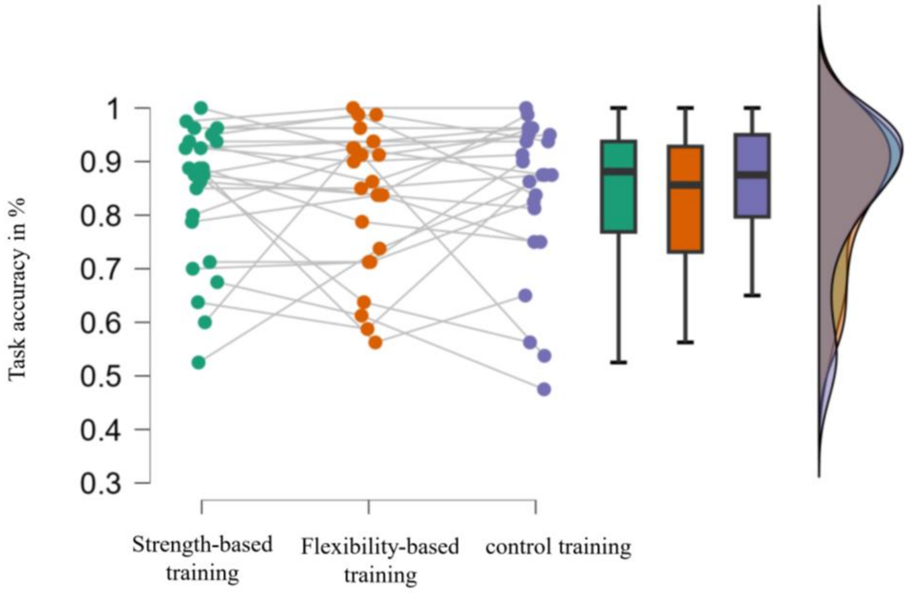


Figure 2

Task Accuracy in Incongruent Trials

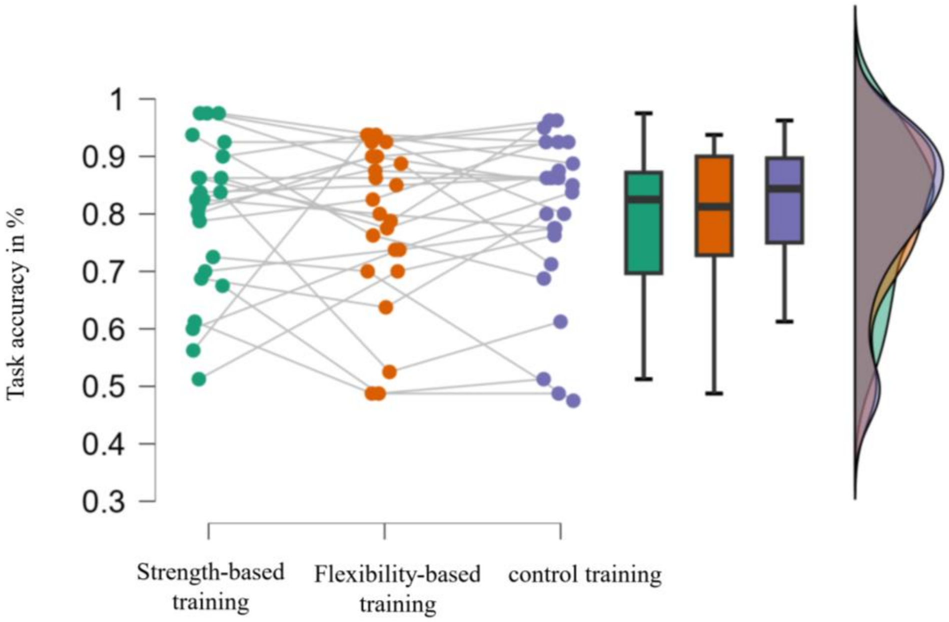


Table 6

Reaction Time, Task Accuracy, and Inverse Efficiency Score (IES)

	Reaction time in ms				Task accuracy in %				Inverse efficiency score in ms			
	congruent		incongruent		congruent		incongruent		congruent		incongruent	
	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
StrT	519.3	91.2	534.0	86.4	83.91	13.1	79.48	13.35	645.58	221.28	701.43	230.17
FlexT	516.2	91.2	529.4	76.8	83.18	13.35	78.75	14.02	642.49	182.23	697.97	188.05
ConT	517.1	96.3	538.9	98.8	83.96	14.8	79.53	14.67	635.49	171.03	700.38	190.25

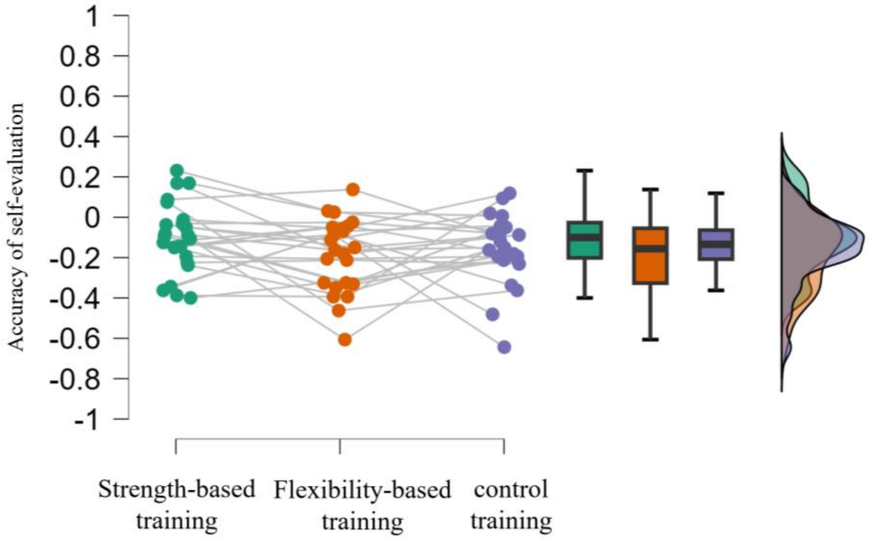
Effect of Physical Exercise on Self-evaluation

Effects of Exercise on Self-evaluation of Task Accuracy

We tested whether training affected the accuracy of self-evaluation when evaluating task accuracy in an Erikson Flanker Task. Figure 3 illustrates accuracy in self-evaluation across the three conditions measured as the difference between self-evaluation and objective task accuracy. Positive values indicate overestimation while negative values indicate underestimation. We found no evidence for a difference in self-evaluation of task accuracy following strength-based, flexibility-based, and the control training, as indicated by a repeated measures ANOVA: $F(2, 46) = 1.672$, p =.199; $\eta_p{}^2$= .068.

Figure 3

Differences between Self-Evaluation and "Objective" Task Accuracy following Strength-Based, Flexibility-Based, and Control Training



Effects of Exercise on Self-evaluation of Reaction Time

We tested whether strength-based and flexibility-based training led to an increase in self-evaluation of their reaction time in an Eriksen Flanker Task. For this, we asked participants to rate how fast they thought they responded to the stimuli during the Erikson Flanker Task after training, using the options 'slower than before training', 'as fast as before the training' and 'faster than before the training'. We compared these responses to objectively measured reaction time. Table 7 shows children's accuracy of self-evaluation regarding their reaction time. The self-evaluations following strength-based and flexibility based training show similar patterns. Fifty percent of the participants overestimated their reactions times. They rated their reactions as faster than they were. Approximately fifteen percent underestimated their reaction times. They rated their reaction times slower than they actually were. While one quarter of the participants rated their reaction times correctly after the strength-based training, only one eighth rated their reaction times correctly after the flexibility-based training. We observe a different pattern of self-evaluations following the control training. Here, approximately half of the participants correctly self-evaluated their reaction times. One quarter each underestimated and overestimated their reaction time.

Table 7

Number of Correct, Overestimated, Underestimated Ratings and No Ratings

	Correctly rated	overestimated	underestimated	no rating
Strength-based training	6 (25%)	12 (50%)	3 (12.5%)	3 (12.5%)
Flexibility-based training	3 (12.5%)	12 (50%)	4 (16.7%)	5 (20.8%)
Control training	11 (45.8%)	6 (25%)	6 (25%)	1 (4.2%)

Effects of Exercise on Emotional Response

Children's Emotional Response immediately after Training

Children's emotional responses to StrT, FlexT, and ConT were assessed as feelings of pleasure and arousal. Differences in pleasure and arousal before and immediately after training were calculated and compared across the training conditions. Tables 8 and 9 present descriptive data on children's feelings of pleasure and arousal. No differences in pleasure between StrT, FlexT, and ConT were evident, as indicated by a Friedman test: χ^2 (2) = 2.65, p = .266; Kendall's W = .063. No differences in arousal between StrT, FlexT, and ConT were evident, as indicated by a Friedman test: χ^2 (2) = 4.29, p = .117; Kendall's W = 0.102. We found no evidence for alterations in children's emotional states across the training conditions.

Table 8

Descriptive Data of Children's Feelings of Pleasure Before and After Training

	pleasure				differences in pleasure	
	pre		post			
	M	*SD*	*M*	*SD*	*M*	*SD*
Strength-based training	7.90	1.18	7.67	2.33	-0.24	2.39
Flexibility based training	7.43	1.69	7.71	2.47	0.29	2.85
Control training	8.00	1.34	8.62	0.74	0.62	1.12

Note: Feelings of pleasure were assessed using Self-Assessment Manikins (Bradley & Lang, 1994) on a scale from 1 = low pleasure to 9 = high pleasure. The differences between pre- and post-measurements were calculated as 'post-measurement minus pre-measurement', *n* = 21, *M* = mean, *SD* = standard deviation.

Table 9

Descriptive Data of Children's Feelings of Arousal before and after Training

	arousal				Differences in arousal	
	pre		post			
	M	*SD*	*M*	*SD*	*M*	*SD*
Strength-based training	4.00	2.93	5.19	3.04	1.19	2.50
Flexibility-based training	4.14	2.59	3.57	2.86	-0.57	3.28
Control training	4.90	2.72	4.62	2.78	-0.29	2.76

Note: Feelings of arousal were assessed using Self-Assessment Manikins (Bradley & Lang, 1994) on a scale from 1 = low arousal to 9 = high arousal. The differences between pre- and post-measurement were calculated as 'post-measurement minus pre-measurement'. *n* = 21, *M* = mean, *SD* = standard deviation.

Parents' Ratings of Children's Emotional Response after Training

We assessed parents' ratings of their children's emotional response to the training. We measured parents' ratings of children's state after StrT, FlexT, and ConT for the remainder of the day. We calculated the difference between these ratings and parents' ratings of children's "usual" emotional state. Tabels 10 and 11 show descriptive data of parents' ratings of children's feelings of pleasure and arousal. No differences between StrT, FlexT, and ConT in parents' ratings of their children's feelings of pleasure were evident, as indicated by a Friedman test: χ^2 (2) = 0.027, p = .987; Kendall's W = 0.001. No differences between StrT, FlexT, and ConT in parents' ratings of their children's feelings of arousal were evident, as indicated by a Friedman test: χ^2(2) = 0.338, p = .845, Kendall's W = 0.007. We did not find evidence for alterations in children's emotional states after the training conditions as rated by their parents.

Table 10

Descriptive Data on Parents' Ratings of their Children's' Feelings of Pleasure

	pleasure				difference in pleasure	
	usual		post			
	M	*SD*	*M*	*SD*	*M*	*SD*
Strength-based training	6.52	1.58	7.29	1.23	0.83	1.70
Flexibility-based training	6.60	1.39	7.08	1.51	0.67	1.90
Control training	6.22	1.13	6.63	1.51	0.41	1.66

Note: Parents' ratings of their children's feelings of pleasure were assessed on a scale from 1 = low pleasure to 9 = high pleasure. The differences between usual perception and post-measurement were calculated as 'post-measurement minus usual perception'. n = 23, M = mean, SD = standard deviation.

Table 11

Descriptive Data on Parents' Ratings of their Children's' Feelings of Arousal

	arousal				differences in arousal	
	usual		post			
	M	*SD*	*M*	*SD*	*M*	*SD*
Strength-based training	6.15	1.67	5.09	1.95	- 1.07	2.07
Flexibility-based training	5.98	1.56	5.28	1.75	- 0.70	2.23
Control training	6.20	1.42	5.00	2.02	- 1.20	1.96

Note: Parents' ratings of their children's feelings of arousal were assessed on a scale from 1 = low arousal to 9 = high arousal. The differences between usual perception and post-measurement were calculated as 'post-measurement minus usual perception'. *n* = 23, *M* = mean, *SD* = standard deviation.

Discussion

Exercise-Type and Attention

In this study, we tested whether strength- and flexibility-based exercise acutely impacts attention in children with ADHD. Strength- and flexibility-based training effectively induced a physiological response. After both training sessions, an increase in heart rate was observed compared to the control condition. The effect was more pronounced after strength-based training. This was in line with our hypothesis. The effect persisted until the subsequent execution of the Erikson Flanker Task, but only after the strength-based training. Despite this distinct physiological response, no evidence for exercise-induced effects on attention was observed, including effects on reaction time, task accuracy, variability of reaction time, and the inverse efficiency score (IES), which accounts for speed-accuracy trade-offs. The speed-accuracy trade-off describes the phenomenon in which reaction times slow when choosing to respond more accurately and accuracy decreases when choosing to respond faster (e.g., Standage et al., 2014). Results of this study contrast previous findings on endurance-based exercise. In those studies, exercise led to improvements in attention in children with ADHD (Ludyga et al., 2020; Medina et al., 2010; Piepmeier et al., 2015; Pontifex et al., 2013). Our findings might be explained in several ways. First, strength-based and flexibility-based training likely induce different physiological responses compared to endurance-based training and to each other; for example, IGF-2 in strength-based training and lactate in endurance-based training (Armstrong & Van Mechelen, 2017). Despite these differences, endurance-based training likely provides a prolonged, more effective training stimulus over an extended period of time. Presumably, this leads to a greater physiological response accompanied by the release of more neurochemicals and a persistently increased blood flow. Strength-based training, on the other hand, is characterized by intermittent training loads and rest phases. Alterations in systemic blood flow are less distinct and remain largely confined locally to the muscle group involved in the specific exercise. Flexibility-based training focuses on improving range of motion and mobility. In this type of training, even less muscular strength and cardiovascular activity are involved. Based on our results, we conclude that strength- and flexibility-based exercise do not induce a physiological response sufficient to alter cognitive function, as studies on endurance-based training have shown. The results of this study indicate that cognitive constraints in ADHD may not be acutely and directly altered through strength- and flexibility-based physical exercise. A direct comparison of all endurance-based, strength-based, and flexibility-based physical exercise might provide insight the validity of our conclusion. Secondly, our results might also be explained by the intensity of training. Training intensity influences potential exercise-induced effects on attention, with moderate intensities yielding the most beneficial effects (e.g., Chang et al., 2015). In our study, subjectively perceived training intensity was generally low to moderate, with a mean applied effort of 4.73 during strength-based training and 3.35 during flexibility-based training, on a scale from 1 to 10. More intense training, e.g., with more repetitions or shorter rest phases, might be needed to induce effects similar to those seen after endurance-based training. Based on our results, we suggest that flexibility-based training is not intense enough to induce effects on attentional parameters. The applied effort in this type of training was relatively low. Increasing intensity in flexibility-based training primarily involves increasing the range of motion and movement angles, which presumably does not lead to an increase in factors associated with improved cognitive functioning (e.g. Audiffren & André, 2019; Solanto, 2002). We argue that higher intensities are required in strength-based training than in endurance-based training to elicit similar effects on cognition. To clarify how training type and training intensity contribute to acute exercise-induced effects on attention, future studies could systematically vary both factors. Beyond these considerations, we would like to point out that while no evidence for positive effects of the two exercise types studied on cognition was observed, physical exercise did not negatively influence attention in children with ADHD either. This is especially important because children expressed less enjoyment during flexibility-based training compared to the control condition, and negative emotionality can be associated with reduced task engagement (Sonuga-Barke, 2003).

Self-Evaluation

In this study, we tested whether physical exercise affects the accuracy of self-evaluation of task accuracy and reaction time in an Erikson Flanker Task in children with ADHD. We hypothesized that physical exercise enhances the accuracy of self-evaluation through improved cue utilization of task-relevant information as a result of enhanced attentional performance. We assumed that improvements in attentional performance modify the 'positive illusory bias,' a phenomenon describing the tendency of children with ADHD to overestimate their competencies (e.g., Hoza et al., 2004; McQuade et al., 2017). We expected greater effects from strength-based training compared to flexibility-based training. To test this hypothesis, we compared the accuracy of self-evaluation across these training types and the control condition. Because the study results revealed no evidence of increased attentional performance, an increase in accuracy of self-evaluation is not expected. Accordingly, no evidence for more accurate self-evaluation of task accuracy or reaction time after strength-based or flexibility-based training was found in this study. To determine whether constraints in attention are indeed a cause of inaccuracy of self-evaluation, it is necessary to sufficiently induce a physiological response that improves attentional performance. Studies on endurance-based training have already shown sufficient changes in attentional performance following physical exercise (e.g., Medina et al., 2010; Pontifex et al., 2013). These studies could be replicated and extended by the incorporating self-evaluation measures. Results might be able to answer our initial research question. It should be noted that this study applied trial-by-trial feedback, meaning that children received information about their performance after each trial. This information may enhance the accuracy of self-evaluating one's own performance. We argue that even though children receive the information about their performance, they tend to misjudge it because this information is not adequately processed due to attentional constraints in ADHD. The results also reveal some unexpected findings. The analysis of self-evaluation of task accuracy does not replicate a clear 'positive illusory bias,' as observed in other studies (e.g., Hoza et al., 2004). Instead, it appears that most children underestimate their task accuracy, as indicated negative values in the difference. Unexpectedly, the analysis of self-evaluation of reaction time reveals a high proportion of overestimation after both training sessions, but not after the control condition. This suggests a decline in the accuracy of self-evaluation after physical exercise, but only for reaction time. This outcome is not easily interpretable. One reason for these results may be methodological in nature. It seems possible that children use other information and cues when self-evaluating their performance than those used to objectively measure performance. In our study, we (non-systematically) observed that children sometimes used extreme values (i.e., minima or maxima) from the trial-to-trial feedback as a cue for their self-evaluation of reaction time. We observed this when children verbalized their self-evaluation (e.g., 'I had 217ms this time and 300ms before the training, so I was quicker this time.'). From a child's perspective, extreme values are valid indicators of their performance. A child may not take all the trials of a cognitive experiment into account and instinctively calculate an overall mean value, but instead consider - similar to a sports competition, such as long jumping - the 'best' try. However, the objective measurement of reaction time was based on average values rather than extreme ones. This suggests that self-evaluation and objective measurement may not have been aligned. The congruence of self-evaluation and performance assessment is implausible if the 'objective' measurement and the subjective evaluation are based on distinct information. This has important implications because it means that children are not necessarily unable to accurately evaluate their performance. Instead, inaccuracies in self-evaluations may result from a lack of alignment between the subjective evaluation and the objective measurement. It seems necessary – for future studies and every-day activities - to inform children which performance indicator is used as an objective measurement so that subjective evaluation and objective measurement can be aligned. An important research question arises from our points of discussion. It seems particularly important to identify which information and cues children actually use to determine how well they have accomplished a task. It is necessary to determine whether this is the same information that objective measurements, parents, and teachers use, or if there is a discrepancy. It is equally important to examine whether there is a difference between children with and without ADHD. Understanding which information and cues children actually use has important implications. It not only helps to identify underlying mechanisms behind the phenomenon of the 'positive illusory bias,' but also allows teachers to foster more accurate self-evaluation. If teachers are aware of how their students use of information, they may either adapt their own evaluation process or support children in using more predictive performance cues.

Emotional response

In this study, we compared the differences in children's self-reported emotional state before and after a strength-based, flexibility-based training session and the control condition in which they watched an animal documentary. We also compared parents' ratings of their children's emotional response throughout the rest of the day with their usual perception. Pleasure and arousal were analysed. Contrary to our hypothesis, we found no evidence for differences in pleasure or arousal between the training sessions and the control session, as reported by the children. This also contrasts with the findings of the study by (Bigelow et al., 2021). Although their exercise regimen, which included 10 minutes of aerobic exercise, was shorter than ours, it was slightly more intense than our exercise (as measured by heart rate). Interestingly, we observed relatively high levels of pleasure across all conditions, with mean scores of 7.9 in the strength-based training, 7.43 in the flexibility-based training, and 8.00 in the control condition. The scale ranged from 1 ('low pleasure') to 9 ('high pleasure'). Ceiling effects may have played a role here. There was also no evidence for differences between the training sessions and the control condition in parents' ratings of their children's feelings of pleasure. We found no evidence for difference between the training sessions and the control condition in children's feelings of arousal and parents' ratings of arousal. This is also contrary to our hypothesis, and may be explained by the intensity level of the training sessions. More intense training may have elicited greater effects on emotional response. Interestingly, there seemed to be an incongruity between children's and parents' ratings. Notably, arousal seemed to be rated higher by parents than by children. There are several possible explanations for discrepancy. Children were asked to rate their emotional state immediately after the training session. The results were compared to their emotional state before the training session and control condition. Parents were asked to rate their children's emotional response for the rest of the day, and results were compared to a 'usual' day. The results indicated methodological differences in measuring emotional response. This finding may also indicate potential time-delayed effects of training on emotional responses. Effects on emotional states might become more pronounced after a prolonged period, indicating the need for methodological adaptation in the timing of measurements in future studies.

Exercise and Academic Achievement

In this study, we tested whether strength-based and flexibility-based training improve attention, self-evaluation, and emotional responses in children with ADHD. We did not find evidence to support our hypothesis that strength-training has a larger effect on the parameters investigated. We assume that the long-term effects of physical exercise play a greater role in improving cognitive performance and emotional responses than acute effects. We also assume that exercise-induced effects on cognitive and metacognitive abilities, as well as emotional responses, are intertwined (see Figure 1). This interplay needs to be investigated further to understand how physical exercise can improve academic performance in ADHD. We encourage future studies to shed light on how cognitive and metacognitive abilities, as well as emotional responses, are linked and to investigate how these factors can improve academic performance in children with and without ADHD.

Strengths and Limitations

To our knowledge, this study is the first to investigate the acute exercise-induced effects of strength-based and flexibility-based physical exercise on attention, self-evaluation, and emotional response in children with ADHD. It contributes to existing findings from studies investigating endurance-based training (e.g., Neudecker et al., 2019). The methodological approach allows for the comparison of cognitive, metacognitive, and emotional effects between sessions within participants rather than between groups. Based on theoretical considerations, we argued that that self-evaluation might become more accurate as a result of exercise-induced effects on cognition. We tested this main hypothesis with the aim of bridging the gap to real-life skills that are essential in scholastic learning environments. The limitations of this study could be addressed in future research. First, our main hypothesis was based on the assumption that physiological responses following physical exercise lead to enhanced attention. For reasons of methodological appropriateness and a different primary research focus, fine-grained physiological responses were not measured in this study. Nevertheless, the underlying physiological processes remain unclear and would likely provide a better understanding of exercise-induced effects in children with and without ADHD. A myriad of biological components likely underlie cognitive enhancement through physical exercise, and future research needs to capture these processes in their complexity. Second, the results of our study remain limited to a short time frame following physical exercise. Children’s attentional, metacognitive and emotional responses were assessed within 10 to 15 minutes after exercising. However, previous research has shown that exercise-induced effects on cognition are time-dependent (e.g., Lambourne & Tomporowski, 2010). Studies that systematically vary the timing and duration of assessments could shed light on the interactions between exercise type and time-delayed effects. Third, this study was conducted to align with previous studies that reported the highest effects of endurance-based exercise at moderate intensities. Based on our findings, which showed no evidence for an effect on attention, metacognition, and emotional response at moderate intensities, we would like to encourage other researchers to investigate whether higher intensities are required in strength-based training compared to endurance-based training to induce effects, or whether strength-based training does not acutely impact attention per se. Fourth, medication to treat ADHD is widely used and impacts brain metabolism. In this study, eleven out of 24 children were taking medication. They either refrained from taking their medication on testing days, or testing was scheduled to minimise the effects of medication. Parents were asked to choose a testing time when their children's medication had worn off. However, we cannot guarantee that the medication had completely worn off. This may have influences the effects of physical exercise. Nevertheless, it is important to investigate effects of physical exercise in children who take regular medication, as physical exercise may serve as a complementary treatment approach.

Statements and Declarations

Data Availability

The data supporting the findings presented in this study are fully available upon request from the corresponding author.

Funding

This study received no specific funding.

Conflict of Interest

The authors declare no financial or non-financial conflicts of interests.

Ethics Approval

The conduct of this study involving human participants was in accordance with the ethical standards set forth in the 2013 Declaration of Helsinki, the ethical guidelines of the German Psychological Society (DGPs), and the institutional ethical standards of the University of Cologne. The conduct of this study was approved by the Ethics Committee of the Faculty of Human Sciences at the University of Cologne (Identification number: SEHF0164).

Authors' Contribution

The authors agree with the content of this work and give their consent to its submission. The authors made the following contributions: SE: conception, study design, training development, recruitment, data acquisition, data analysis, data interpretation, writing; SF: recruitment, manuscript revision; EA: supervision; JG: conception, resources, manuscript revision, and supervision.

Acknowledgements

We would like to thank Armin Oster for his support in developing and filming the training sessions, as well as for his professional advice. We are also grateful to thank Britta Moers for her professional proofreading, Selina Faist for her recruitment efforts, and Simon Gavanda for his valuable input throughout the conceptualisation of the study.

References

Armstrong, N., & Van Mechelen, W. (2017). *Oxford Textbook of Children's Sport and Exercise Medicine* Oxford University Press.

Arnold, L. E., Hodgkins, P., Kahle, J., Madhoo, M., & Kewley, G. (2020). Long-term outcomes of ADHD: academic achievement and performance. *Journal of Attention Disorders*, *24*(1), 73-85.

Audiffren, M., & André, N. (2019). The exercise–cognition relationship: A virtuous circle. *Journal of Sport and Health Science*, *8*(4), 339-347.

Barry, T. D., Lyman, R. D., & Klinger, L. G. (2002). Academic underachievement and attention-deficit/hyperactivity disorder: The negative impact of symptom severity on school performance. *Journal of School Psychology*, *40*(3), 259-283.

Basso, J. C., & Suzuki, W. A. (2017). The effects of acute exercise on mood, cognition, neurophysiology, and neurochemical pathways: A review. *Brain Plasticity*, *2*(2), 127-152.

Bigelow, H., Gottlieb, M. D., Ogrodnik, M., Graham, J. D., & Fenesi, B. (2021). The differential impact of acute exercise and mindfulness meditation on executive functioning and psycho-emotional well-being in children and youth with ADHD. *Frontiers in Psychology*, *12*, 660845.

BKK Gesundheit.Online BMI. [Calculator for childrens' BMI]. Retrieved from: www.bkk-gesundheit.de/ernaehrung/bmi-rechner-fuer-kinder; February 2024.

Boes, K. (2017). Handbuch Motorische Tests. [Handbook motoric testing]. Hogrefe, Göttingen

Bradley, M. M., & Lang, P. J. (1994). Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, *25*(1), 49-59.

Buesch, D., Utesch, T., & Marschall, F. (2021). Development and evaluation of the "Anstrengungsskala Sport"(Effort Scale Sport). *German Journal of Exercise and Sport Research*, 1-6.

Bunford, N., Evans, S. W., & Wymbs, F. (2015). ADHD and emotion dysregulation among children and adolescents. *Clinical child and family psychology review*, *18*, 185-217.

Castelli, D. M., Hillman, C. H., Buck, S. M., & Erwin, H. E. (2007). Physical fitness and academic achievement in third-and fifth-grade students. *Journal of Sport and Exercise Psychology*, *29*(2), 239-252.

Christiansen, H., Hirsch, O., Albrecht, B., & Chavanon, M.-L. (2019). Attention-deficit/hyperactivity disorder (ADHD) and emotion regulation over the life span. *Current Psychiatry Reports*, *21*, 1-11.

Christiansen, L., Beck, M. M., Bilenberg, N., Wienecke, J., Astrup, A., & Lundbye-Jensen, J. (2019). Effects of exercise on cognitive performance in children and adolescents with ADHD: potential mechanisms and evidence-based recommendations. *Journal of clinical medicine*, *8*(6), 841.

Caterino, M. C., & Polak, E. D. (1999). Effects of two types of activity on the performance of second-, third-, and fourth-grade students on a test of concentration. *Perceptual and Motor Skills*, *89*(1), 245-248.

Chaddock, L., Hillman, C. H., Pontifex, M. B., Johnson, C. R., Raine, L. B., & Kramer, A. F. (2012). Childhood aerobic fitness predicts cognitive performance one year later. *Journal of Sports Sciences*, *30*(5), 421-430.

Chang, Y.-K., Chu, C.-H., Wang, C.-C., Wang, Y.-C., Song, T.-F., Tsai, C.-L., & Etnier, J. L. (2015). Dose–response relation between exercise duration and cognition. *Medicine & Science in Sports & Exercise, 47*(1), 159-165.

Chang, Y.-K., Liu, S., Yu, H.-H., & Lee, Y.-H. (2012). Effect of acute exercise on executive function in children with attention deficit hyperactivity disorder. *Archives of Clinical Neuropsychology*, *27*(2), 225-237.

Coghill, D. R., Seth, S., Pedroso, S., Usala, T., Currie, J., & Gagliano, A. (2014). Effects of methylphenidate on cognitive functions in children and adolescents with attention-deficit/hyperactivity disorder: evidence from a systematic review and a meta-analysis. *Biological Psychiatry, 76*(8), 603-615.

Daley, D., & Birchwood, J. (2010). ADHD and academic performance: why does ADHD impact on academic performance and what can be done to support ADHD children in the classroom? *Child: Care, Health and Development*, *36*(4), 455-464.

Daley, D., Van der Oord, S., Ferrin, M., Danckaerts, M., Doepfner, M., Cortese, S., Group, E. A. G. (2014). Behavioral interventions in attention-deficit/hyperactivity disorder: a meta-analysis of randomized controlled trials across multiple outcome domains. *Journal of the American Academy of Child & Adolescent Psychiatry*, *53*(8), 835-847. e835.

De Greeff, J. W., Bosker, R. J., Oosterlaan, J., Visscher, C., & Hartman, E. (2018). Effects of physical activity on executive functions, attention and academic performance in preadolescent children: a meta-analysis. *Journal of Science and Medicine in Sport, 21*(5), 501-507.

Den Heijer, A. E., Groen, Y., Tucha, L., Fuermaier, A. B., Koerts, J., Lange, K. W., . . . Tucha, O. (2017). Sweat it out? The effects of physical exercise on cognition and behavior in children and adults with ADHD: a systematic literature review. *Journal of Neural Transmission*, *124*, 3-26.

DuPaul, G. J., Weyandt, L. L., & Janusis, G. M. (2011). ADHD in the classroom: Effective intervention strategies. *Theory into Practice, 50*(1), 35-42.

Emeh, C. C., Mikami, A. Y., & Teachman, B. A. (2018). Explicit and implicit positive illusory bias in children with ADHD. *Journal of Attention Disorders, 22*(10), 994-1001.

Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & Psychophysics, 16*(1), 143-149.

Faigenbaum, A. D., Kraemer, W. J., Blimkie, C. J., Jeffreys, I., Micheli, L. J., Nitka, M., & Rowland, T. W. (2009). Youth resistance training: updated position statement paper from the national strength and conditioning association. *The Journal of Strength & Conditioning Research, 23,* S60-S79.

Faraone, S. V., Banaschewski, T., Coghill, D., Zheng, Y., Biederman, J., Bellgrove, M. A., . . . Manor, I. (2021). The world federation of ADHD international consensus statement: 208 evidence-based conclusions about the disorder. *Neuroscience & Biobehavioral Reviews, 128,* 789-818.

Frazier, T. W., Youngstrom, E. A., Glutting, J. J., & Watkins, M. W. (2007). ADHD and achievement: Meta-analysis of the child, adolescent, and adult literatures and a concomitant study with college students. *Journal of Learning Disabilities, 40*(1), 49-65.

Fritz, K. M., & O'Connor, P. J. (2016). Acute exercise improves mood and motivation in young men with ADHD symptoms. *Med Sci Sports Exerc, 48*(6), 1153-1160.

Grassmann, V., Alves, M. V., Santos-Galduroz, R. F., & Galduróz, J. C. F. (2017). Possible cognitive benefits of acute physical exercise in children with ADHD: a systematic review. *Journal of Attention Disorders, 21*(5), 367-371.

Graziano, P. A., & Garcia, A. (2016). Attention-deficit hyperactivity disorder and children's emotion dysregulation: A meta-analysis. *Clinical Psychology Review, 46,* 106-123.

Herold, F., Müller, P., Gronwald, T., & Müller, N. G. (2019). Dose–response matters!–a perspective on the exercise prescription in exercise–cognition research. *Frontiers in Psychology, 10,* 2338.

Hillman, C. H., Kamijo, K., & Scudder, M. (2011). A review of chronic and acute physical activity participation on neuroelectric measures of brain health and cognition during childhood. *Preventive Medicine, 52,* S21-S28.

Hoza, B., Gerdes, A. C., Hinshaw, S. P., Arnold, L. E., Pelham Jr, W. E., Molina, B. S., . . . Hechtman, L. (2004). Self-perceptions of competence in children with ADHD and comparison children. *Journal of Consulting and Clinical Psychology, 72*(3), 382.

Hoza, B., Pelham Jr, W. E., Dobbs, J., Owens, J. S., & Pillow, D. R. (2002). Do boys with attention-deficit/hyperactivity disorder have positive illusory self-concepts? *Journal of Abnormal Psychology, 111*(2), 268.

IBM Corp. (2021). IBM SPSS Statistics for Windows, Version 29.0 [Computer Software]. IBM Corp.

JASP Team (2024). JASP (Version 0.18.3) [Computer software].

Jangmo, A., Stålhandske, A., Chang, Z., Chen, Q., Almqvist, C., Feldman, I., . . . Kuja-Halkola, R. (2019). Attention-deficit/hyperactivity disorder, school performance, and effect of medication. *Journal of the American Academy of Child & Adolescent Psychiatry*, *58*(4), 423-432.

Kent, K. M., Pelham Jr, W. E., Molina, B. S., Sibley, M. H., Waschbusch, D. A., Yu, J., . . . Karch, K. M. (2011). The academic experience of male high school students with ADHD. *Journal of Abnormal Child Psychology*, *39*(3), 451-462.

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.

Kortekaas-Rijlaarsdam, A. F., Luman, M., Sonuga-Barke, E., & Oosterlaan, J. (2019). Does methylphenidate improve academic performance? A systematic review and meta-analysis. *European Child & Adolescent Psychiatry*, *28*, 155-164.

Kramer, A. F., Erickson, K. I., & Colcombe, S. J. (2006). Exercise, cognition, and the aging brain. *Journal of applied physiology*, *101*(4), 1237-1242.

Lambez, B., Harwood-Gross, A., Golumbic, E. Z., & Rassovsky, Y. (2020). Non-pharmacological interventions for cognitive difficulties in ADHD: A systematic review and meta-analysis. *Journal of Psychiatric Research*, *120*, 40-55.

Lambourne, K., & Tomporowski, P. (2010). The effect of exercise-induced arousal on cognitive task performance: a meta-regression analysis. *Brain Research, 1341*, 12-24.

Langberg, J. M., Dvorsky, M. R., Molitor, S. J., Bourchtein, E., Eddy, L. D., Smith, Z., . . . Evans, S. W. (2016). Longitudinal evaluation of the importance of homework assignment completion for the academic performance of middle school students with ADHD. *Journal of school psychology, 55*, 27-38.

Lenhard, W. & Lenhard, A. (2022). Calculation of effect sizes. Retrieved from: https://www.psychometrica.de/effektstaerke.html. Psychometrica. DOI: 10.13140/RG.2.2.17823.92329

Liang, X., Qiu, H., Wang, P., & Sit, C. H. (2022). The impacts of a combined exercise on executive function in children with ADHD: A randomized controlled trial. *Scandinavian Journal of Medicine & Science in Sports, 32*(8), 1297-1312.

Loe, I. M., & Feldman, H. M. (2007). Academic and educational outcomes of children with ADHD. *Journal of Pediatric Psychology, 32*(6), 643-654.

Ludyga, S., Brand, S., Gerber, M., Weber, P., Brotzmann, M., Habibifar, F., & Pühse, U. (2017). An event-related potential investigation of the acute effects of aerobic and coordinative exercise on inhibitory control in children with ADHD. *Developmental cognitive neuroscience*, *28*, 21-28.

Ludyga, S., Gerber, M., Mücke, M., Brand, S., Weber, P., Brotzmann, M., & Pühse, U. (2020). The acute effects of aerobic exercise on cognitive flexibility and task-related heart rate variability in children with ADHD and healthy controls. *Journal of attention disorders, 24*(5), 693-703.

Massetti, G. M., Lahey, B. B., Pelham, W. E., Loney, J., Ehrhardt, A., Lee, S. S., & Kipp, H. (2008). Academic achievement over 8 years among children who met modified criteria for attention-deficit/hyperactivity disorder at 4–6 years of age. *Journal of Abnormal Child Psychology, 36*, 399-410.

McQuade, J. D., Mendoza, S. A., Larsen, K. L., & Breaux, R. P. (2017). The nature of social positive illusory bias: Reflection of social impairment, self-protective motivation, or poor executive functioning? *Journal of Abnormal Child Psychology, 45*, 289-300.

Medina, J. A., Netto, T. L., Muszkat, M., Medina, A. C., Botter, D., Orbetelli, R., . . . Miranda, M. C. (2010). Exercise impact on sustained attention of ADHD children, methylphenidate effects. *ADHD Attention Deficit and Hyperactivity Disorders, 2*, 49-58.

Meßler, C. F., Holmberg, H.-C., & Sperlich, B. (2018). Multimodal therapy involving high-intensity interval training improves the physical fitness, motor skills, social behavior, and quality of life of boys with ADHD: a randomized controlled study. *Journal of Attention Disorders, 22*(8), 806-812.

Metcalfe, J., & Finn, B. (2008). Evidence that judgments of learning are causally related to study choice. *Psychonomic Bulletin & Review, 15*(1), 174-179.

Moreau, D., Kirk, I. J., & Waldie, K. E. (2017). High-intensity training enhances executive function in children in a randomized, placebo-controlled trial. *Elife, 6*, e25062.

Mulser, L., & Moreau, D. (2023). Effect of acute cardiovascular exercise on cerebral blood flow: A systematic review. *Brain Research*, 148355.

Neudecker, C., Mewes, N., Reimers, A. K., & Woll, A. (2019). Exercise interventions in children and adolescents with ADHD: a systematic review. *Journal of Attention Disorders, 23*(4), 307-324.

Ng, Q. X., Ho, C. Y. X., Chan, H. W., Yong, B. Z. J., & Yeo, W.-S. (2017). Managing childhood and adolescent attention-deficit/hyperactivity disorder (ADHD) with exercise: A systematic review. *Complementary Therapies in Medicine, 34*, 123-128.

Owens, J. S., Goldfine, M. E., Evangelista, N. M., Hoza, B., & Kaiser, N. M. (2007). A critical review of self-perceptions and the positive illusory bias in children with ADHD. *Clinical Child and Family Psychology Review, 10*, 335-351.

Piepmeier, A. T., Shih, C.-H., Whedon, M., Williams, L. M., Davis, M. E., Henning, D. A., . . . Etnier, J. L. (2015). The effect of acute exercise on cognitive performance in children with and without ADHD. *Journal of Sport and Health Science, 4*(1), 97-104.

Pontifex, M. B., Saliba, B. J., Raine, L. B., Picchietti, D. L., & Hillman, C. H. (2013). Exercise improves behavioral, neurocognitive, and scholastic performance in children with attention-deficit/hyperactivity disorder. *The Journal of Pediatrics*, *162*(3), 543-551.

Robinson, K., Riley, N., Owen, K., Drew, R., Mavilidi, M. F., Hillman, C. H., . . . Lubans, D. R. (2023). Effects of Resistance Training on Academic Outcomes in School-Aged Youth: A Systematic Review and Meta-Analysis. *Sports Medicine*, 1-15.

Salari, N., Ghasemi, H., Abdoli, N., Rahmani, A., Shiri, M. H., Hashemian, A. H., ... & Mohammadi, M. (2023). The global prevalence of ADHD in children and adolescents: a systematic review and meta-analysis. *Italian Journal of Pediatrics*, *49*(1), 48.

Sedgwick, J. A., Merwood, A., & Asherson, P. (2019). The positive aspects of attention deficit hyperactivity disorder: a qualitative investigation of successful adults with ADHD. *ADHD Attention Deficit and Hyperactivity Disorders*, *11*, 241-253.

Singh, A. S., Saliasi, E., Van Den Berg, V., Uijtdewilligen, L., De Groot, R. H., Jolles, J., . . . Diamond, A. (2019). Effects of physical activity interventions on cognitive and academic performance in children and adolescents: a novel combination of a systematic review and recommendations from an expert panel. *British Journal of Sports Medicine*, *53*(10), 640-647.

Skriver, K., Roig, M., Lundbye-Jensen, J., Pingel, J., Helge, J. W., Kiens, B., & Nielsen, J. B. (2014). Acute exercise improves motor memory: exploring potential biomarkers. *Neurobiology of Learning and Memory*, *116*, 46-58.

Solanto, M. V. (2002). Dopamine dysfunction in AD/HD: integrating clinical and basic neuroscience research. *Behavioural Brain Research*, *130*(1-2), 65-71.

Sonuga-Barke, E. J. (2002). Psychological heterogeneity in AD/HD—a dual pathway model of behaviour and cognition. *Behavioural Brain Research*, *130*(1-2), 29-36.

Sonuga-Barke, E. J. (2003). The dual pathway model of AD/HD: an elaboration of neuro-developmental characteristics. *Neuroscience & Biobehavioral Reviews*, *27*(7), 593-604.

Standage, D., Blohm, G., & Dorris, M. C. (2014). On the neural implementation of the speed-accuracy trade-off. *Frontiers in Neuroscience, 8*, 236.

Tomporowski, P. D. (2003a). Cognitive and behavioral responses to acute exercise in youths: A review. *Pediatric Exercise Science*, *15*(4), 348-359.

Tomporowski, P. D. (2003b). Effects of acute bouts of exercise on cognition. *Acta psychologica*, *112*(3), 297-324.

Tomporowski, P. D., McCullick, B., Pendleton, D. M., & Pesce, C. (2015). Exercise and children's cognition: The role of exercise characteristics and a place for metacognition. *Journal of Sport and Health Science, 4*(1), 47-55.

Van Hall, G., Stømstad, M., Rasmussen, P., Jans, Ø., Zaar, M., Gam, C., . . . Nielsen, H. B. (2009). Blood lactate is an important energy source for the human brain. *Journal of Cerebral Blood Flow & Metabolism*, *29*(6), 1121-1129.

Vysniauske, R., Verburgh, L., Oosterlaan, J., & Molendijk, M. L. (2020). The effects of physical exercise on functional outcomes in the treatment of ADHD: a meta-analysis. *Journal of attention disorders, 24*(5), 644-654.

Volz-Sidiropoulou, E., Boecker, M., & Gauggel, S. (2016). The positive illusory bias in children and adolescents with ADHD: further evidence. *Journal of Attention Disorders, 20*(2), 178-186.

World Medical Association. (2001). Declaration of Helsinki. Ethical principles for medical research involving human subjects. *Bulletin of the World Health Organization*, *79*(4), 373.

Yeung, M. K., Lee, T. L., & Chan, A. S. (2020). Neurocognitive development of flanker and Stroop interference control: A near-infrared spectroscopy study. *Brain and Cognition*, *143*, 105585.

4 General Discussion

The primary aim of this dissertation was to examine the potential effects of skill practice, prompts, and physical exercise on the accuracy of self-evaluation. Three studies were conducted. Table 2 provides an overview of the studies included in this dissertation and the methods used to investigate potential effects. Each study used a distinct approach that may lead to an increase in the accuracy of self-evaluation.

Study I examined the impact of practicing the skill to be evaluated. The study was conducted with 167 eighth-grade students. School students practiced concept mapping – concept map construction or concept map study – and evaluated their concept mapping skills.

Study II examined the impact of prompting, i.e., explicit instruction in text reading. In this study, 162 pre-service biology teachers were prompted to ask themselves global comprehension questions while reading a biology text. Resource- and deficit-oriented questions were used. The study was conducted as an online study. Pre-service teachers evaluated their text comprehension before completing a short comprehension test.

Study III examined the effects of physical exercise on the accuracy of self-evaluation. An improvement in attentional performance after exercise was suggested to enhance the accuracy of self-evaluation. Strength-based and flexibility-based physical exercises were investigated. Twenty-four children with ADHD participated in this study. Children evaluated their task accuracy and reaction time after completing an Eriksen Flanker Task. The findings of these three studies will be discussed in terms of their impact on the accuracy of self-evaluation, methodological approaches, and future directions.

Table 2

Overview of Studies Included in this Dissertation

	Study I	**Study II**	**Study III**
Research Object	Skill practice	Prompting	Physical exercise
Sample	Eighth-grade students (*N* = 167)	Pre-service biology teachers (*N* = 162)	Children with ADHD (*N* = 24)
Methodological Design	Field intervention study controlled, quasi-experimental, between-subjects, two-factor	Online short-term study controlled, randomised, between-subjects, one-factor	Laboratory study controlled, within-subjects, repeated-measures
Self-Evaluated Ability	Concept mapping skills	Text comprehension	Attention: task accuracy and reaction time
Measurement of Congruence	Correlation	Differences	Differences and categorisation
Main Findings	Slight increased alignment of self-evaluation after concept map construction training compared to concept map study training	No evidence for an effect of resource-oriented or deficit-oriented self-questioning	No evidence for an effect of physical exercise on attention, hence, no evidence for an effect of physical exercise on self-evaluation

4.1 Discussion and Future Research

4.1.1 Discussion and Future Research – Study I

Study I examined the effects of concept map training on the accuracy of students' self-evaluations. Students completed either concept map construction training or concept map study training sessions. Students rated their concept-mapping skills in a subsequent learning phase. The ratings of their concept-mapping skills were compared to an "objective" assessment of concept map quality. The findings of the study indicated a slight advantage of concept map construction training over concept map study training. The Spearman correlation revealed a slightly higher effect size for concept map construction training ($r_s = .66$, $p < 0.001$, $n = 30$) compared to concept map study training ($r_s = .52$, $p = 0.004$, $n = 28$). This finding is in line with the findings of Kruger and Dunning, (1999). In their study, they showed that participants in the bottom quartile of a skill level particularly overestimated their skills in humour, grammar, and logic. However, training the skill itself, i.e., logical reasoning, led to an improvement in the accuracy of self-evaluation. The authors describe this outcome as an "increase in calibration" that occurs as a result of training. In this study, it was not differentiated between low- and high-skilled participants but investigated overall alignment through correlations. A slight increase in the effect size, indicating a slightly improved alignment after concept map construction training, was observed. The results are also in line with the findings of Schroeder et al. (2018), who found that effects of constructing concept maps was generally higher than studying concept maps.[8]

The results of the Study I may be explained by the differences between concept map construction and concept map study training. Concept map construction is a more active form of training, in which learners practice creating concept maps themselves. Skeleton concept maps were completed, propositions were compiled, arrow directions were added, and concept maps were evaluated. Concept map study training is a more passive form of training, in which the "reading" of concept maps was encouraged. Learners used worksheets. They identified key concepts and relationships in concept maps and determined reading directions. The level of procedural knowledge was higher in concept map construction training than in concept map study training. This outcome suggests that improving the accuracy of self-evaluation may be achieved through enhancing the skill to be evaluated itself. Interestingly, this outcome suggests that self-evaluation itself does not need to be addressed specifically. Instead, the

[8] Note that this meta-analysis investigated the effects of concept maps construction and concept map study primarily on learning outcomes, rather than on self-evaluation.

accuracy of self-evaluation occurs as a by-product of skill development. Whether this is true for all learners under all conditions, still needs to be investigated.

Importantly, in the present study we did not investigate whether a Kruger-Dunning effects was evident as shown by others (Kruger & Dunning, 1999). Only a few studies have investigated whether the Kruger-Dunning effect is present in biology education. These studies have been carried out with university students (e.g., Osterhage et al., 2019; Ziegler & Montplaisir, 2014). Less is known about the Kruger-Dunning effect in biology education.

Several future research questions arise from our study results. Two study ideas emerge from our study results as research gaps that need to be filled. First, future studies could address the question of whether practicing the skill itself is superior to practicing self-evaluation. Both approaches may be addressed within one study, also including a combination of skill and self-evaluation practice. The outcome of such a study might provide important implications for teachers and educators. They could provide insights into what to focus on during teaching: skill practice, self-evaluation practice, or a combination of both when intending to increase the accuracy of self-evaluation. Second, the existence of the Kruger-Dunning effect may be investigated in biology education in school students. An understanding of the baseline accuracy of self-evaluations in biology education provides important insights into improving the accuracy.

4.1.2 Discussion and Future Research – Study II

Study II examined the effects of prompting during text reading on the accuracy of self-evaluation. In an online learning environment, pre-service teachers were asked to read a biology text. Pre-service teachers were either prompted to use a resource-oriented question, a deficit-oriented question, or they were not prompted to use any question while reading. The resource-oriented question was: “What have I already understood?” The deficit-oriented question was: “What have I not yet understood?” Both questions were global comprehension questions. After reading the text, pre-service teachers evaluated their text comprehension. Subsequently, they completed a comprehension test. Discrepancies between self-evaluated and “objectively” measured text comprehension were calculated and compared across the three conditions. The findings of our study did not support the hypothesis that resource-oriented prompts enhance the accuracy of self-evaluation of comprehension after text reading.

Our finding contrasts with other studies that have shown that relatively simple learning strategies improve the accuracy of self-evaluation, such as summarizing, re-reading, or retrieval practice (Miller & Geraci, 2014; Rawson et al., 2000). However, it needs to be noted that the present study initiated immediate recall through self-questioning, not delayed recall. Self-evaluation was also assessed immediately after the reading phase. Other studies have shown that delaying the retrieval of keywords or self-evaluation improves the accuracy of self-evaluation compared to an immediate recall (Nelson & Dunlosky, 1991; Thiede et al., 2003; Thiede et al., 2005). Delaying both the recall of subject knowledge and self-evaluation assessment may be more beneficial for improving the accuracy of self-evaluation. Moreover, the specificity of the question may play an important role. Other authors have shown that specific prompts enhance the accuracy of self-evaluation compared to global prompts (Kramarski & Kohen, 2017).

Beyond these considerations, another aspect arises that may explain why prompting in our study did not enhance the accuracy of self-evaluation: the level of processing. To explain the study findings, this dissertation applies the levels-of-processing framework (Craik & Lockhart, 1972). This framework suggests a hierarchical order of processing, ranging from “shallow” to “deep” processing (e.g., Craik, 2002).[9] Although this framework was developed within cognitive psychology, it may also be applicable to self-questioning: a

[9] “Deep” processing describes the analyses of “meaning, inference, and implication.” Shallow processing describes the analyses of “surface form, colour, loudness, and brightness.” (Craik, 2002, p. 308). The description refers to stimuli typically used in laboratory experiments. The transfer of the terms “shallow” and “deep” is still to be established in learning settings.

comprehension-related self-question may be processed deeply or shallowly, i.e., understood and further processed. Answers to a self-question may be elaborate and specific if processing is deep. Conversely, answers may be less elaborate and unspecific if processing is shallow. Naturally, the level of processing is likely to affect how the information generated during the answering process is used as a cue for further self-evaluation. An answer resulting from deep processing is likely to have a positive impact on the accuracy of self-evaluation, if only because more specific information is generated and actively kept in a state of availability; see embedded processes model of working memory (e.g., Cowan, 1999). An answer resulting from shallow processing is likely to have a low impact on the accuracy of self-evaluation. In the present study, the level of processing was difficult to control. Pre-service teachers may have taken notes to answer the question, indicating deep processing, but they may also have ignored the question and moved on to reading the next part of the text, indicating shallow processing. In an online setting, controlling the level of processing is more difficult than in an in-person setting. Factors such as a given time frame for answering the question and additional instructional aid are easier to regulate.

Future studies may investigate several research questions arising from these two considerations. First, there is a need for a clear definition of what “deep” and “shallow” processing mean in learning settings. The concept of “level of processing” has been related to information processing in cognitive psychology. However, there is a need to describe how this knowledge relates to learning settings and how these types of processing can actually be measured. If we know how to measure “deep” and “shallow” processing, we actually may be able to compare online and in-person learning settings. Potential differences in the processing of prompts in online and in-person learning settings may be investigated.

Second, the role of time may be of particular importance for the efficacy of prompts. In this study, the immediate effects of prompting during text reading on an immediate assessment of self-evaluation were investigated. However, effects may not always be visible immediately. Long-term effects of repeated self-questioning on the accuracy of self-evaluation may be of interest.

4.1.3 Discussion and Future Research – Study III

Study III examined the effects of physical exercise on self-evaluation in children with ADHD. It was argued that an increase in the accuracy of self-evaluation results from improved attentional performance. Self-evaluation of task accuracy (i.e., proportion of correct responses) and reaction time were measured. The differences between self-evaluation and "objective" task accuracy were compared across the three conditions (strength-based physical exercise vs. flexibility-based physical exercise vs. control condition). Self-evaluations of reaction time were analysed through a categorization, and the results were described. Study findings did not support our initial hypothesis. They did not reveal an effect of physical exercise on attention like previous studies did (e.g., Piepmeier et al., 2015).

Following the initial argumentation, improvement in self-evaluation cannot be expected based on this first finding. If attention is not altered, self-evaluation will not be altered (at least not due to changes in attention). Indeed, a positive effect on self-evaluation was not found in this study. Task accuracy and reaction time were not positively impacted by physical exercise. Two unexpected findings were observed in this study. First, the analyses of task accuracy has not replicated a clear "positive illusory bias", a phenomenon describing overly positive self-evaluations in children with ADHD found in other studies (Chan & Martinussen, 2016; Hoza et al., 2002; Owens & Hoza, 2003). Instead of overestimating, they rather underestimated their performance. This was observed for the task accuracy. Second, study findings revealed a high proportion of overestimation of reaction time after both training sessions that were not as evident in the control condition.

Methodological imprecisions may of course have played a role (see Chapter 4.2). However, there are other possible explanations. The less pronounced positive illusory bias will be addressed. One reason that the illusory bias was less evident in the evaluation of task accuracy may be the use of feedback. Feedback was given trial-by-trial. Because children received information on every trial they completed, the accuracy of self-evaluation may have been improved. However, we did not observe in increased alignment. Instead it seemed that the accuracy was still impaired but towards underestimation. This was not tested, but descriptively observed through the index that was used to compare the three conditions. Explanations for this can only be speculative. The mechanisms behind the self-illusory bias are still widely unknown (e.g., Crisci et al., 2022; Owens et al., 2007). In this study, we assume that cognitive impairments are a major cause for overly positive self-evaluations. However, other mechanisms have been proposed as well, such as social impairments, self-protective motivation, or language skills (Crisci et al., 2022; McQuade et al., 2017).

Additionally, there is tentative evidence that more severe symptoms of hyperactivity-impulsivity are associated with the positive illusory bias but not more severe symptoms of inattention (Owens & Hoza, 2003). In the present study, the ADHD subtypes were not captured. Of course, this finding by Owens and Hoza (2003) questions the role of attention as a cause for overly positive self-evaluation. It remains to be answered of what actually causes the overly positive self-evaluations if cognitive functions seem less causal. It may be argued that motivational and emotional aspects are also relevant (see: the dual pathway model of ADHD; Sonuga-Barke, 2003). These aspects may explain a greater proportion than previously expected.

Future studies may investigate the impact of emotions on self-evaluation in children with ADHD. For example, emotions may be deliberately manipulated through situations that stimulate pleasure and frustration (to an extent not beyond daily life experiences). The impact on self-evaluation can be examined subsequently.

4.1.4 Overall Discussion and Future Research

Self-evaluation is considered an inherent part of self-regulated learning (Zimmerman, 1990). Self-evaluations occur in the self-reflection phase of the cyclical process of self-evaluation, in which own abilities and performance are judged (Zimmerman, 1990). Empirical evidence suggests that self-evaluations influence subsequent learning behaviour (Metcalfe & Finn, 2008). For learners to adopt adequate learning behaviour, self-evaluations need to be accurate; however they do not necessarily align with “objective” performance indicators. This dissertation examined distinct approaches potentially affecting the accuracy of self-evaluation. It examined the potential effects of skill practice (Study I), prompting (Study II), and physical exercise (Study III) on self-evaluation. In Study I, tentative evidence of an effect of skill practice on self-evaluation was found. In this study with eighth-grade students, self-evaluation and “objective” performance were slightly more aligned after concept map construction training than after concept map study training. Study II found no evidence of an effect of prompting during text reading in an online learning setting with pre-service teachers was found. Neither resource-oriented nor deficit-oriented prompting appeared to affect the accuracy of self-evaluation. Study III found no effect of physical exercise, i.e., strength-based and flexibility-based training, on attention. It was hypothesised that changes in attention would affect self-evaluation. No evidence was found for a positive impact of physical exercise on self-evaluation.

The results of Study I support findings from other studies that have shown an increase in the accuracy of self-evaluation through the development of the skill itself (Kruger & Dunning, 1999). These results raise the question of whether specific training to foster the accuracy of self-evaluation is indeed needed. A recent study has shown that specific training in monitoring one’s own performance increased the accuracy of self-evaluation (Händel et al., 2020). Interestingly, the effects observed through metacognitive monitoring training exceeded those observed after repeated testing[10], which increased the skill itself (Händel et al., 2020). Nevertheless, applying repeated testing without specific metacognitive practice improves the accuracy of self-evaluation (Naujoks et al., 2022). There is a need to understand how skill development, on the one hand, and specific training in metacognitive skills, on the other hand, impact the accuracy of self-evaluation – both separately and in combination.

At the same time, several studies have shown that metacognitive strategies are not universally effective. For instance, only approximately half of the university students in an

[10] The testing effect is a widely examined phenomenon describing positive effects on retention through testing, e.g., Rowland (2014)

introductory biology course who were prompted to use metacognitive skills actually followed through with their plans (Stanton et al., 2015). In another study, metacognitive training itself appeared to be effective only if cognitive training was ineffective (Leopold & Leutner, 2015). It has also been suggested that only the regular use of prompts in combination with appropriate in-depth cognitive strategies, such as note-taking, may benefit the learning process (Moser et al., 2017). The use of global evaluation criteria results in less accurate self-evaluations, as observed in learners with ADHD (Prevatt et al., 2012). Study II contributes to this body of literature by providing no evidence of an effect of prompting (i.e., global resource-oriented and deficit-oriented prompts) on the accuracy of self-evaluation in an online learning setting.

Other studies have shown that the accuracy of self-evaluation can be improved. For instance, judgment training and self-testing have been shown to enhance the accuracy of self-evaluation (Händel et al., 2020; Naujoks et al., 2022). Additionally, delaying judgments of learning and aligning judgment of learning items with test questions improve the accuracy of self-evaluation (Rhodes & Tauber, 2011). The use of specific rating criteria improved the accuracy of self-evaluation in learners with ADHD (Prevatt et al., 2012). The results of the studies included in this dissertation further highlight the complexity of self-evaluation and its accuracy. Despite its complexity, implications for improving the accuracy of self-evaluation will be drawn from the results of the studies in this dissertation and the theoretical background as well as empirical evidence from previous research (see Chapter 4.3 of this dissertation). The next paragraph discusses the theoretical background applied in this dissertation.

The three studies included in this dissertation used approaches from cognitive psychology as their theoretical background and as a basis for explaining the study results. For example, the concept of metacognition (Flavell, 1979) was applied in Studies I and II, the embedded processes model (Cowan, 1999) in Study II, the cue-utilization framework (Koriat, 1997) in Studies II and III, and the “neurotrophic hypothesis” (Audiffren & André, 2019) in Study III. The outcomes of the studies in this dissertation suggest that these cognitive backgrounds may not be sufficient to fully explain effects on self-evaluation. Particularly, emotions might play a vital role in the formation of self-evaluation but are not typically integrated in theoretical models of self-evaluation. The role of emotions has not been a primary focus of this dissertation, mainly due to its scope limitations. Emotional aspects were measured in each study (e.g., enjoyment of concept mapping in Study I, contentment with their learning in Study II, and pleasure and arousal in Study III) but these variables were not specifically analysed regarding their impact on self-evaluation. Nevertheless, an understanding of involved emotions might be crucial for gaining a deeper understanding of the formation of self-evaluation. The impact of emotions could be examined by manipulating[11] emotional states such as pleasure and frustration. Potential effects on self-evaluation could then be tested.

One critical question that has accompanied the process of writing this dissertation is: Is it truly necessary to accurately self-evaluate one's own performance and abilities to be a self-regulated learner, and to gain an understanding of the content that interests us? A meta-analysis of 95 interventions indeed showed that self-evaluation is associated with academic performance (Donker et al., 2014). However, learners who underestimate their own abilities and performance – as shown in studies investigating the Big-Fish-Little-Pond effect (Marsh, 1987) or in the evaluation of participants with skills in the upper quartile (Kruger & Dunning, 1999) - may even benefit from of their inaccurate self-evaluations. The benefits of underestimation may lie in a subjectively perceived, continuous need to improve abilities and skills, which may lead to exceptionally high performance. If the sole or most important argument for the use of self-evaluation is to improve performance and academic outcomes, then we would not strive for accuracy of self-evaluation but for underestimation. Certainly, there are other arguments for the necessity to improve accuracy of self-evaluation, such as the importance of self-evaluation as a social function or as an indicator and supporter of well-being (see, for example, McQuade et al., 2016). Although accurately self-evaluate one's own ability and performance may not be essential for comprehension and learning performance - particularly in cases of underestimation – it certainly supports a better understanding of oneself.

[11] The term "manipulation" refers to the deliberate variation of the independent variable. Only this manipulation enables the investigation of a cause-and-effect relationship. A study design that influences emotions in human participants is subject to review by an ethics committee. In such a study, emotional states may only be influenced to an extent that does not go beyond everyday experiences.

4.2 Methodological Assessment of Self-Evaluation

This dissertation addresses the accuracy of self-evaluation that is understood as the congruence between subjective evaluations of one's own learning and "objectively" measured indicators of ability or performance. Each study within this dissertation took a different methodological approach to determining the accuracy of self-evaluations. Each method has its advantages and disadvantages.

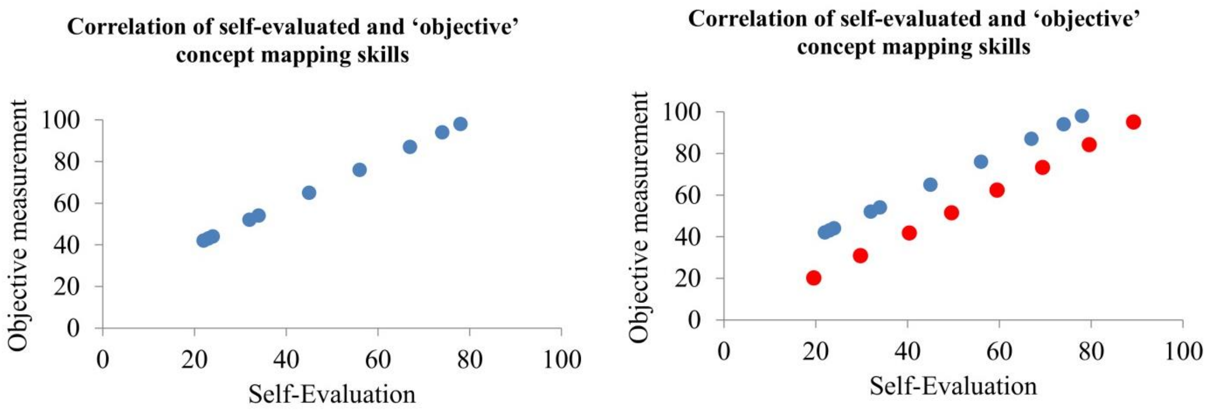
Study I investigated the accuracy of self-evaluation in concept mapping skills. For this purpose, self-evaluations and concept map quality were correlated using Spearman's rank correlations. Self-evaluation was measured using five statements, such as "I paid attention to the direction of the arrows" and "I labelled all the arrows," rated on a three-step emoticon-based scale (Elzen-Rump & Leutner, 2007). The "objective" assessment of concept mapping skills was conducted using a scoring system suggested by Clausen and Christian (2012). Each proposition was assigned a score based on the type of relation, labels, and connecting structures; e.g., 0 = two linked concepts without substantial relation, 5 = cause-effect relation with labelled arrow. Using correlations to determine the accuracy of self-evaluation has the major advantage of being easy to understand and interpret, as correlations represent a basic methodological approach. However, complete congruence may not be captured using this method. Spearman's rank correlation is a special form of the Pearson correlation, which is calculated using the covariance of two variables:

$$cov\ (x, y) = \frac{1}{n} \cdot \sum (x\ - \bar{x}\ ) \ \cdot (y - \bar{y})$$

This covariance is used to determine the Spearman correlation coefficient. To do this, the covariance is divided by the product of the standard deviations of both variables, x and y. For a Spearman correlation, ranks are used instead of the raw data. Correlations measure both the direction and the strength of a linear relationship. They reflect how closely the data points follow an ideal straight line. A perfect linear relationship corresponds to a correlation coefficient of r = +1 when an increase in one variable is completely associated an increase in another. However, a correlation of r = +1 does not necessarily indicate true congruency. This can be illustrated with an example. Figure 5 shows two scatterplots of fictional data describing the relationship of self-assessed and "objective" measurements of concept mapping skills.

Figure 5

Schematic Correlations of Self-Evaluated and "Objective" Concept Mapping Skills

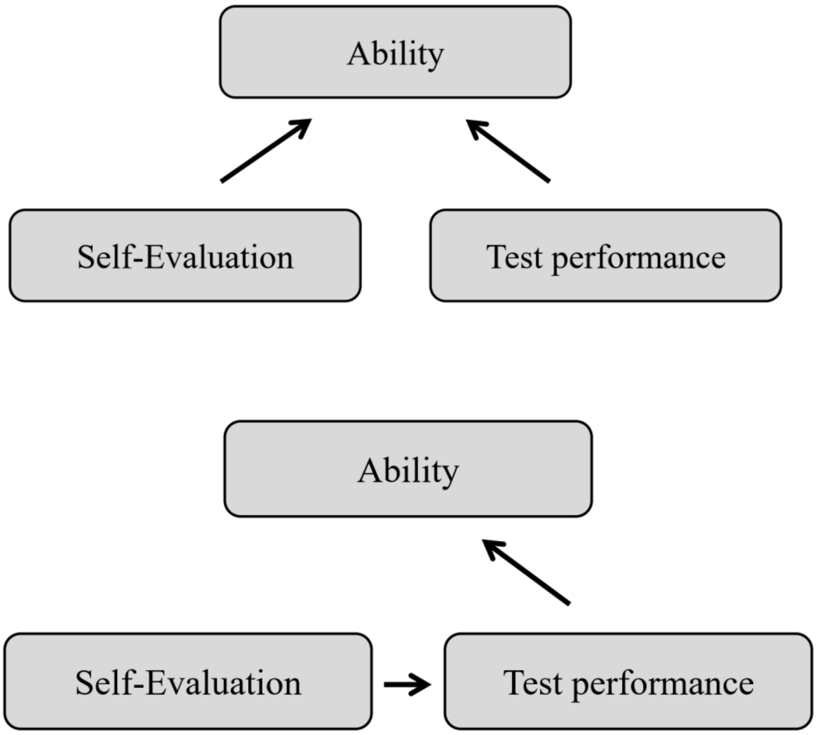


Both the self-evaluation and "objective" measurement apply a scale from 0 to 100. Calculating the correlation coefficient would result in a value of $r = +1$. This might be interpreted as congruency. However, this interpretation overlooks a potential systematic bias. Such a potential bias becomes apparent when we add the red trend line in the figure on the right. The red dots represent true congruency, where a self-assessed value of 20 corresponds to a value of 20 of the "objective" scale (30 = 30, 40 = 40, and so on). The blue dots, on the other hand, are systematically shifted. A self-evaluated score of 20 corresponds to an "objective" score of 40, a score of 40 corresponds to 60, and so on. This means every data point is shifted 20 above the line that represents true congruence between subjective and "objective" measurement. This means participants systematically underestimated their concept mapping skills. To complicate matters further, the self-evaluated and "objective" measurements did not use a common 0-100 scale but two differently scaled instruments, making a potential systematic shift harder to detect. To conclude, correlations may not reveal true congruence, and alternative methodological approaches may be needed.

Study II investigated the accuracy of self-evaluation regarding the text comprehension of a biology text. For this, the numerical difference between self-evaluation and "objectively" measured text comprehension was calculated. Self-evaluation was measured using a visual analogue scale ranging from 0 to 100 %, in response to the question: “How much of the content did you comprehend?” Learning performance was calculated as the proportion of correct answers to test questions. The difference between self-evaluated and “objective” learning performance was compared across three conditions (resource-oriented question, deficit-oriented question, no question). The advantage of this approach is that it allows to measure true congruence which could not be done through correlations. The calculation is fairly easy to understand, and mathematical operations are even simpler than those used in correlations. Using this approach in Study II highlights the importance of the study design. A general overestimation was observed, which may be strongly connected to the time of assessment. Self-evaluation was assessed before the comprehension test was administered. The participants did not yet have any information about the upcoming test. Information, such as the difficulty level of individual test items, was not yet available. If the assessment had been conducted after the test, more accurate self-evaluations would likely have resulted. However, this was not intended. The aim was not to assess participants’ self-evaluation of their test performance, but rather of their text comprehension. This is an important distinction. See Figure 6 for this. These thoughts also points to the “objectivity” of performance assessment. This will be addressed later in this chapter.

Figure 6

The Conceptual Relationship between Ability, Self-Evaluation and Test Performance



Study III investigated the accuracy of self-evaluation regarding task accuracy and reaction time in children with ADHD. For this, two self-assessment scales were used. First, self-evaluation of task accuracy was measured using a five-step, thumb-based scale. Second, self-evaluation of reaction time was assessed using a three-step scale, comparing reaction time after the training sessions to that before the training. Congruence between self-evaluation and “objective” task accuracy was assessed by calculating the difference between the two. This was similar to the approach taken in Study II. However, the approach in Study III required converting the self-evaluation scale into a percentage scale. This transformation may have led to a loss of accuracy, for example, in comparison to a visual analogue scale. However, measuring self-evaluation using the thumb-based scale appeared to be an appropriate method for assessing children's self-evaluation. Congruence between self-evaluation and "objective" reaction time measurements was assessed using a categorization system. As a prerequisite, the standard deviation was used to define threshold for over- and underestimation. This approach was useful and produced an output that was easy to understand. However, the calculation steps are not easy to follow, and it depended on a social criterion.

Importantly, it should to be noted that complete "objectivity" of learning performance measurement is rarely present. Indicators of learning performance themselves often depend on estimation by instruments or individuals. In this dissertation, "objective" performance was measured in multiple ways. In Study I, concept-mapping skills were measured by three independent raters using a rating system to estimate concept map quality. The interrater agreement was Fleiss'κ = .82 (map 3), Fleiss'κ = .96 (map 1), and Fleiss'κ = 1 (map 2). In Study II, learning performance was measured using a learning test. This test showed an interrater agreement, based on intraclass estimates for the open-ended question, ranging from .66 to .96. In Study III, response accuracy and reaction time were measured using E-Prime 3.0 Software. It becomes evident that "objective" measurements are never entirely objective.[12] These instruments and measurements are themselves a source of error. This is particularly important when the supposed "objective" measurement relies on subjective estimation, as it typically does in learning settings. For example, teacher ratings show great variance when evaluating the performance of children with ADHD (Langberg et al., 2008).

Because of this, the quality of the measuring instruments needs to be ensured, as well as the alignment between subjective evaluation and "objective" measurement scales, so that both measurement are able to capture what they intend to capture.

[12] Because of constraints in securing complete objectivity, quotation marks are used throughout the entire dissertation for the terms "objectivity" and "objective".

4.3 Implications and Relevance to Biology Education

A shift towards active learning and the use of metacognitive strategies has been proposed in science and biology education (American Association for the Advancement of Science, 2011; Tanner, 2012). These suggestions build on an overarching aim in education to support life-long learning (Taranto & Buchanan, 2020). Self-regulated learning is understood as a framework that provides a conceptual orientation to support learners in becoming active participants in their own learning process, with self-evaluation being an integral part of most theoretical approaches to self-regulated learning (e.g., Panadero, 2017). Although self-regulated learning has been shown to be a valuable concept to support learning, the use of specific strategies remains somewhat unclear in many areas, including science education (Zohar & Barzilai, 2013). However, the use of these specific strategies may be of particular importance to biology because the subject itself with its complex phenomena, methodological approaches, and technical terms demands supportive strategies for learning to be able to understand subject content (see Chapter 1.2 of this dissertation: Metacgonition and Self-Evaluation in Biology Education). Moreover, ineffective use of metacognitive strategies and inaccurate self-evaluations have been shown to be as present in this subject as in many others (Osterhage et al., 2019).

The dissertation approaches self-evaluation in biology education from an interdisciplinary point of view. It highlights the importance – if not the necessity – of multiple approaches to improve the accuracy of self-evaluation in biology education. It may not have produced step-by-step instructions for improving the accuracy of self-evaluation in biology teaching; nevertheless, it offers an interdisciplinary perspective on the complexity of the accuracy of self-evaluation that may serve as a starting point for better understanding within domain-specific applications of metacognitive instructions. This dissertation applies concepts of cognitive and educational psychology to biology education. It strengthens our understanding of self-evaluation, including declarative knowledge about the formation of self-evaluation and individual cue-use. It also points to methodological considerations in the assessment of self-evaluation, for instance, the alignment of subjective and "objective" measurements used to measure abilities and performance. Moreover, this dissertation included fundamental research by including a study that focused on attention, a prerequisite for the use of metacognitive strategies (see Study III).

The decision to include a study investigating self-evaluation in children with ADHD was based on multiple considerations. First, by investigating a group that is particularly prone to inaccurate self-evaluations, I hoped to better understand how the accuracy of self-evaluation can be improved for many learners, including those experiencing learning barriers. Second, ADHD is a fairly frequently diagnosed developmental disorder (Salari et al., 2023). Teachers are likely to encounter students with ADHD and observe negative impact on academic achievement (Arnold et al., 2020), highlighting the need for effective support strategies. Third, biology education is a school subject that can support dialogue about psychological disorders such as ADHD. The Standing Conference of the Ministers of Education and Cultural Affairs ("Kultusministerkonferenz"; KMK, 2020) explicitly states that biology education should foster respect and responsibility for other living beings and one's own health. In this sense, teaching about ADHD in biology classes – similar to anorexia nervosa or alcoholism – can be an essential part of the subject. Including the topic of ADHD in biology classes may help children better understand the disorder, develop empathy for one another, and thus improve the quality of life for those affected. A broader societal understanding – primarily supported through education – may increase the well-being of many who are affected by the disorder. Moreover, teaching about ADHD can easily be integrated into biological content such as brain structure, hormonal regulation, and neurotransmitter systems. These reasons support the inclusion of ADHD as a topic in biology education. This dissertation does not directly implement these ideas into practice. However, it aims to serve as a starting point for further discussion of these ideas.

Overall, this dissertation contributes to our understanding of self-evaluation in biology education, including in children with ADHD, who are particularly prone to inaccurate self-evaluations. This dissertation highlights the combined use of cognitive and metacognitive strategies (Study I), the importance of using specific prompts rather than global prompts (Study II), and the necessity of aligning subjective and "objective" measurements (Study I to III).

4.4 The Role of Non-Significant Research Results

This chapter addresses two issues that are important to all areas of science that work with statistical analyses: the interpretation of non-significant results and the file drawer problem (Dienes, 2014; Rosenthal, 1979). These issues arise not only because the studies in this dissertation yielded mostly non-significant results, but also because misinterpretations of their meaning are common (Hemming et al., 2022).

The file drawer problem was already described in the 1970s (Rosenthal, 1979). It refers to a specific type of bias in the publication of research findings (Rosenthal, 1979). The file drawer problem describes the tendency to publish statistically significant results, leading to journals being “filled with the 5% of the studies that show Type I errors, while the file drawers back at the lab are filled with the 95% of the studies that show non-significant (e.g., $p > .05$) results” (Rosenthal, 1979, p. 638). The file drawer problem is similar to the publication bias observed in many areas (e.g., Mesquida et al., 2023). The reasons for such bias may be manifold. Publishers and journals may prefer publishing results with statistical effects, as they are more marketable, and researchers may believe they can only succeed in their field if they produce statistically significant results or are able to “prove”[13] an effect. While these tendencies may not be entirely unfounded, they can have detrimental effects on the overall research process. Studies are not replicated due to methodological constraints (Maxwell et al., 2015), and in some cases researchers may even go so far as to manipulate data (Stroebe et al., 2012). These consequences undermine confidence in science with potentially detrimental societal effects.

Over recent years, efforts have been made to increase the reliability and trustworthiness of scientific findings. For example, guidance on interpreting *p*-values has been published, such as “The ASA statement on p-values: context, process, and purpose” (Wasserstein & Lazar, 2016) and “Valid p-values behave exactly as they should: Some misleading criticisms of p-values and their resolution with s-values” (Greenland, 2019). Similar studies were published regarding the interpretation of non-significant results (Dienes, 2014). The problem with non-significant results is described as follows: “A non-significant result can mean one of two things: either there is evidence for the null-hypothesis and against a theory that predicted a difference (or relationship); or else that the data are insensitive in distinguishing the theory from the null hypothesis and nothing follows from the data at all”

[13] The use of the term “prove” is not appropriate in scientific contexts. Science does not prove hypotheses; science tests hypotheses - or develops hypotheses in qualitative research.

(Dienes et al., 2014, p. 1). One proposed solution – though not without criticism – is the use of Bayes factors (Dienes, 2014).

To conclude, non-significant and actual null-results are of immense importance to the research process. They need to be published and discussed. Correct interpretation of research results requires stronger methodological understanding and process awareness. Enhancing our integrity as researchers is fundamental to improving the quality of research findings.

5 Conclusion

The aim of this dissertation was to examine potential effects on the accuracy of self-evaluation. It aimed to answer the practice-oriented question of how to improve the accuracy of self-evaluation by applying psychological theories. It applied approaches from cognitive and educational psychology to biology education to understand how self-evaluations are formed and how their accuracy can be improved. More broadly, it aimed to foster self-regulated learning in biology education to enable learners to address real-life challenges, both related to biology and beyond. This interdisciplinary dissertation comprises three studies investigating skill practice, prompting, and physical exercise. The overall results show a slight effect of skill practice on the accuracy of self-evaluation, but no evidence for an effect of prompting or physical exercise. The aims of this dissertation may be broader than the answers provided by the individual studies included. Nevertheless, this dissertation may serve as a starting point for future research aiming to better understand ourselves and the complex world we live in.

Literature

American Association for the Advancement of Science (2011). *Vision and Change: A Call to Action.* Final Report. Washington, DC: AAAS.

American Psychiatric Association (2013). *Diagnostic and statistical manual of mental disorders: DSM-5* (Vol. 5). American Psychiatric Association Washington, DC.

Arbuckle, T. Y., & Cuddy, L. L. (1969). Discrimination of item strength at time of presentation. *Journal of Experimental Psychology, 81*(1), 126.

Arnold, L. E., Hodgkins, P., Kahle, J., Madhoo, M., & Kewley, G. (2020). Long-term outcomes of ADHD: academic achievement and performance. *Journal of Attention Disorders*, *24*(1), 73-85.

Baddeley, A. (1992). Working memory. *Science*, *255*(5044), 556-559.

Barry, T. D., Lyman, R. D., & Klinger, L. G. (2002). Academic underachievement and attention-deficit/hyperactivity disorder: The negative impact of symptom severity on school performance. *Journal of School Psychology, 40*(3), 259-283.

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.

Boekaerts, M. (2011). Emotions, emotion regulation, and self-regulation of learning: center for the study of learning and instruction, Leiden University, The Netherlands, and KU Leuven. In *Handbook of Self-regulation of Learning and Performance* (pp. 422-439). Routledge.

Boekaerts, M., & Cascallar, E. (2006). How far have we moved toward the integration of theory and practice in self-regulation? *Educational Psychology Review*, *18*, 199-210.

Boekaerts, M., Maes, S., & Karoly, P. (2005). Self-regulation across domains of applied psychology: Is there an emerging consensus? *Applied Psychology: an International Review*, *54*(2).

Brandstädter, K., Harms, U., & Grossschedl, J. (2012). Assessing system thinking through different concept-mapping practices. *International Journal of Science Education*, *34*(14), 2147-2170.

Cambridge Advanced Learner's Dictionary & Thesaurus (2024). *Meaning of self-evaluation in English.* retrieved December 16, 2024 from https://dictionary.cambridge.org/dictionary/english/self-evaluation

Chan, T., & Martinussen, R. (2016). Positive illusions? The accuracy of academic self-appraisals in adolescents with ADHD. *Journal of Pediatric Psychology, 41*(7), 799-809.

Conner, L. N. (2007). Cueing metacognition to improve researching and essay writing in a final year high school biology class. *Research in Science Education*, *37*, 1-16.

Council Resolution of 27 June 2002 on lifelong learning, 1-3 163 (2002). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32002G0709(01) retrevied January 2025 from https://op.europa.eu/en/publication-detail/-/publication/0bf0f197-5b35-4a97-9612-19674583cb5b

Cowan, N. (1988). Evolving conceptions of memory storage, selective attention, and their mutual constraints within the human information-processing system. *Psychological Bulletin, 104*(2), 163.

Cowan, N. (1999). An embedded-processes model of working memory. *Models of working memory: Mechanisms of active maintenance and executive control*, *20*(506), 1013-1019.

Cowan, N., Morey, C. C., & Naveh-Benjamin, M. (2020). An embedded-processes approach to working memory. *Working Memory: The State of the Science*, *44*.

Craik, F. I. (2002). Levels of processing: Past, present... and future? *Memory*, *10*(5-6), 305-318.

Craik, F. I., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior, 11*(6), 671-684.

Crisci, G., Cardillo, R., & Mammarella, I. C. (2022). The processes underlying positive illusory bias in ADHD: The role of executive functions and pragmatic language skills. *Journal of Attention Disorders, 26*(9), 1245-1256.

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology*, *5*, 781.

Dignath, C., & Mevarech, Z. (2021). Introduction to special issue mind the gap between research and practice in the area of teachers' support of metacognition and SRL. *Metacognition and Learning*, *16*, 517-521.

Dignath, C., & Sprenger, L. (2020). Can you only diagnose what you know? The relation between teachers' self-regulation of learning concepts and their assessment of students' self-regulation. *Frontiers in Education*. Vol. 5, p. 585683. Frontiers Media SA.

Donker, A. S., De Boer, H., Kostons, D., Van Ewijk, C. D., & van der Werf, M. P. (2014). Effectiveness of learning strategy instruction on academic performance: A meta-analysis. *Educational Research Review*, *11*, 1-26.

Dunlosky, J., & Metcalfe, J. (2008). *Metacognition*. Sage Publications.

Dunlosky, J., Mueller, M. L., Morehead, K., Tauber, S. K., Thiede, K. W., & Metcalfe, J. (2021). Why does excellent monitoring accuracy not always produce gains in memory performance? *Zeitschrift für Psychologie*, *229*(2), 104.

Dunlosky, J., & Rawson, K. A. (2012). Overconfidence produces underachievement: Inaccurate self evaluations undermine students' learning and retention. *Learning and Instruction, 22*(4), 271-280.

Dunlosky, J., & Thiede, K. W. (2004). Causes and constraints of the shift-to-easier-materials effect in the control of study. *Memory & Cognition*, *32*, 779-788.

Dunning, D. (2011). The Dunning–Kruger effect: On being ignorant of one's own ignorance. In *Advances in Experimental Social Psychology* (Vol. 44, pp. 247-296). Elsevier.

Dye, K. M., & Stanton, J. D. (2017). Metacognition in upper-division biology students: Awareness does not always lead to control. *CBE—Life Sciences Education*, *16*(2), ar31.

Efklides, A. (2011). Interactions of metacognition with motivation and affect in self-regulated learning: The MASRL model. *Educational Psychologist*, *46*(1), 6-25.

Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & Psychophysics*, *16*(1), 143-149.

Evagorou, M., Korfiatis, K., Nicolaou, C., & Constantinou, C. (2009). An investigation of the potential of interactive simulations for developing system thinking skills in elementary school: A case study with fifth-graders and sixth-graders. *International Journal of Science Education, 31*(5), 655-674.

Evangelista, N. M., Owens, J. S., Golden, C. M., & Pelham, W. E. (2008). The positive illusory bias: do inflated self-perceptions in children with ADHD generalize to perceptions of others? *Journal of Abnormal Child Psychology, 36*, 779-791.

Fang, J., Huang, X., Zhang, M., Huang, F., Li, Z., & Yuan, Q. (2018). The big-fish-little-pond effect on academic self-concept: A meta-analysis. *Frontiers in Psychology*, *9*, 1569.

Faraone, S. V., Banaschewski, T., Coghill, D., Zheng, Y., Biederman, J., Bellgrove, M. A., Newcorn, J.H., Gignac, M, Al Saud, N.M., Manor, I., Rhode, L.A., Yang, L., Cortese, S., Almagor, D., Stein, M.A., Albatti, T.H., Aljoudi, H.D., …(2021). The world federation of ADHD international consensus statement: 208 evidence-based conclusions about the disorder. *Neuroscience & Biobehavioral Reviews*, *128*, 789-818.

Faraone, S. V., Biederman, J., & Mick, E. (2006). The age-dependent decline of attention deficit hyperactivity disorder: a meta-analysis of follow-up studies. *Psychological Medicine*, *36*(2), 159-165.

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

Hacker, D. J., Dunlosky, J., & Graesser, A. C. (2009). A growing sense of "agency". In *Handbook of Metacognition in Education* (pp. 1-4). Routledge.

Hadwin, A., Järvelä, S., & Miller, M. (2017). Self-regulation, co-regulation, and shared regulation in collaborative learning environments. In *Handbook of Self-regulation of Learning and Performance* (pp. 83-106). Routledge.

Händel, M., Harder, B., & Dresel, M. (2020). Enhanced monitoring accuracy and test performance: Incremental effects of judgment training over and above repeated testing. *Learning and Instruction*, *65*, 101245.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of educational research*, *77*(1), 81-112.

Hemming, K., Javid, I., & Taljaard, M. (2022). A review of high impact journals found that misinterpretation of non-statistically significant results from randomized trials was common. *Journal of Clinical Epidemiology*, *145*, 112-120.

Hertzog, C., Dixon, R. A., & Hultsch, D. F. (1990). Relationships between metamemory, memory predictions, and memory task performance in adults. *Psychology and Aging*, *5*(2), 215.

Hoza, B., Gerdes, A. C., Hinshaw, S. P., Arnold, L. E., Pelham Jr, W. E., Molina, B. S., . . . Hechtman, L. (2004). Self-perceptions of competence in children with ADHD and comparison children. *Journal of Consulting and Clinical Psychology*, *72*(3), 382.

Hoza, B., Pelham Jr, W. E., Dobbs, J., Owens, J. S., & Pillow, D. R. (2002). Do boys with attention-deficit/hyperactivity disorder have positive illusory self-concepts? *Journal of Abnormal Psychology*, *111*(2), 268.

Hoza, B., Waschbusch, D. A., Owens, J. S., Pelham, W. E., & Kipp, H. (2001). Academic task persistence of normally achieving ADHD and control boys: Self-evaluations, and attributions. *Journal of Consulting and Clinical Psychology*, *69*(2), 271.

James, W. (1890). The principles of psychology. *Henry Holt*.

Karlen, Y., Hertel, S., & Hirt, C. N. (2020). Teachers' professional competences in self-regulated learning: An approach to integrate teachers' competences as self-regulated learners and as agents of self-regulated learning in a holistic manner. In *Frontiers in Education* (Vol. 5, p. 159). Frontiers Media SA.

Kent, K. M., Pelham Jr, W. E., Molina, B. S., Sibley, M. H., Waschbusch, D. A., Yu, J., . . . Karch, K. M. (2011). The academic experience of male high school students with ADHD. *Journal of Abnormal Child Psychology*, *39*(3), 451-462.

King, J. F., Zechmeister, E. B., & Shaughnessy, J. J. (1980). Judgments of knowing: The influence of retrieval practice. *The American Journal of Psychology*, 329-343.

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.

Koriat, A., & Bjork, R. A. (2005). Illusions of competence in monitoring one's knowledge during study. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*(2), 187.

Kramarski, B., & Kohen, Z. (2017). Promoting preservice teachers' dual self-regulation roles as learners and as teachers: Effects of generic vs. specific prompts. *Metacognition and Learning, 12,* 157-191.

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.

Kultusministerkonferenz (KMK). (2020). Bildungsstandards im Fach Biologie für die Allgemeine Hochschulreife. [Standing Conference of the Ministers of Education and Cultural Affairs in Germany. Scholastic Standards in Biology.] Retrieved September 2024 from https://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2020/2020_06_18-BildungsstandardsAHR_Biologie.pdf.

Lambez, B., Harwood-Gross, A., Golumbic, E. Z., & Rassovsky, Y. (2020). Non-pharmacological interventions for cognitive difficulties in ADHD: A systematic review and meta-analysis. *Journal of Psychiatric Research*, *120*, 40-55.

Langberg, J. M., Dvorsky, M. R., Molitor, S. J., Bourchtein, E., Eddy, L. D., Smith, Z., . . . Evans, S. W. (2016). Longitudinal evaluation of the importance of homework assignment completion for the academic performance of middle school students with ADHD. *Journal of School Psychology*, *55*, 27-38.

Langberg, J. M., Epstein, J. N., Altaye, M., Molina, B. S., Arnold, L. E., & Vitiello, B. (2008). The transition to middle school is associated with changes in the developmental trajectory of ADHD symptomatology in young adolescents with ADHD. *Journal of Clinical Child & Adolescent Psychology*, *37*(3), 651-663.

Leopold, C., & Leutner, D. (2015). Improving students' science text comprehension through metacognitive self-regulation when applying learning strategies. *Metacognition and Learning, 10,* 313-346.

Ludyga, S., Brand, S., Gerber, M., & Pühse, U. (2017). Exercise as neuroenhancer in children with ADHD: Cognitive and behavioural effects. In *Physical Activity and Educational Achievement* (pp. 191-212). Routledge.

Marsh, H. W. (1987). The big-fish-little-pond effect on academic self-concept. *Journal of Educational Psychology, 79*(3), 280.

Marshall, R. M., Hynd, G. W., Handwerk, M. J., & Hall, J. (1997). Academic underachievement in ADHD subtypes. *Journal of Learning Disabilities, 30*(6), 635-642.

Martin, B. L., Mintzes, J. J., & Clavijo, I. E. (2000). Restructuring knowledge in biology: Cognitive processes and metacognitive reflections. *International Journal of Science Education*, *22*(3), 303-323.

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does “failure to replicate” really mean?. *American Psychologist*, *70*(6), 487.

Mazzoni, G., & Cornoldi, C. (1993). Strategies in study time allocation: Why is study time sometimes not effective? *Journal of Experimental Psychology: General*, *122*(1), 47.

McCarthy, K. S., Likens, A. D., Johnson, A. M., Guerrero, T. A., & McNamara, D. S. (2018). Metacognitive overload!: Positive and negative effects of metacognitive prompts in an intelligent tutoring system. *International Journal of Artificial Intelligence in Education*, *28*, 420-438.

McQuade, J. D., Mendoza, S. A., Larsen, K. L., & Breaux, R. P. (2017). The nature of social positive illusory bias: Reflection of social impairment, self-protective motivation, or poor executive functioning? *Journal of Abnormal Child Psychology*, *45*, 289-300.

Mesquida, C., Murphy, J., Lakens, D., & Warne, J. (2023). Publication bias, statistical power and reporting practices in the Journal of Sports Sciences: potential barriers to replicability. *Journal of Sports Sciences*, *41*(16), 1507-1517.

Metcalfe, J. (2002). Is study time allocated selectively to a region of proximal learning? *Journal of Experimental Psychology: General*, *131*(3), 349.

Metcalfe, J. (2009). Metacognitive judgments and control of study. *Current Directions in Psychological Science*, *18*(3), 159-163.

Metcalfe, J., & Finn, B. (2008). Evidence that judgments of learning are causally related to study choice. *Psychonomic Bulletin & Review*, *15*(1), 174-179.

Metcalfe, J., & Kornell, N. (2003). The dynamics of learning and allocation of study time to a region of proximal learning. *Journal of Experimental Psychology: General*, *132*(4), 530.

Miller, T. M., & Geraci, L. (2014). Improving metacognitive accuracy: How failing to retrieve practice items reduces overconfidence. *Consciousness and Cognition*, *29*, 131-140.

Miyake, A., & Shah, P. (1999). *Models of working memory*. Citeseer.

Moser, S., Zumbach, J., & Deibl, I. (2017). The effect of metacognitive training and prompting on learning success in simulation-based physics learning. *Science Education*, *101*(6), 944-967.

Mulser, L., & Moreau, D. (2023). Effect of acute cardiovascular exercise on cerebral blood flow: A systematic review. *Brain Research*, *1809*, 148355.

Naujoks, N., Harder, B., & Händel, M. (2022). Testing pays off twice: Potentials of practice tests and feedback regarding exam performance and judgment accuracy. *Metacognition and Learning*, *17*(2), 479-498.

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In *Psychology of Learning and Motivation* (Vol. 26, pp. 125-173). Elsevier.

Nelson, T. O., & Dunlosky, J. (1991). When people's judgments of learning (JOLs) are extremely accurate at predicting subsequent recall: The "delayed-JOL effect". *Psychological Science*, *2*(4), 267-271.

Nelson, T. O., Gerler, D., & Narens, L. (1984). Accuracy of feeling-of-knowing judgments for predicting perceptual identification and relearning. *Journal of Experimental Psychology: General*, *113*(2), 282.

O'Neill, M. E., & Douglas, V. I. (1991). Study strategies and story recall in attention deficit disorder and reading disability. *Journal of Abnormal Child Psychology*, *19*(6), 671-692.

Osterhage, J. L., Usher, E. L., Douin, T. A., & Bailey, W. M. (2019). Opportunities for self-evaluation increase student calibration in an introductory biology course. *CBE—Life Sciences Education*, *18*(2), ar16.

Owens, J. S., Goldfine, M. E., Evangelista, N. M., Hoza, B., & Kaiser, N. M. (2007). A critical review of self-perceptions and the positive illusory bias in children with ADHD. *Clinical Child and Family Psychology Review*, *10*, 335-351.

Owens, J. S., & Hoza, B. (2003). The role of inattention and hyperactivity/impulsivity in the positive illusory bias. *Journal of Consulting and Clinical Psychology*, *71*(4), 680.

Palennari, M. (2016). Exploring the correlation between metacognition and cognitive retention of students using some biology teaching strategies. *Journal of Baltic Science Education*, *15*(5), 617-629.

Panadero, E. (2017). A review of self-regulated learning: Six models and four directions for research. *Frontiers in Psychology*, *8*, 422.

Peters, E., & Kitsantas, A. (2010). The effect of nature of science metacognitive prompts on science students' content and nature of science knowledge, metacognition, and self-regulatory efficacy. *School Science and Mathematics*, *110*(8), 382-396.

Piepmeier, A. T., Shih, C.-H., Whedon, M., Williams, L. M., Davis, M. E., Henning, D. A.,. Etnier, J. L. (2015). The effect of acute exercise on cognitive performance in children with and without ADHD. *Journal of Sport and Health Science*, *4*(1), 97-104.

Pintrich, P. (2000). The role of goal orientation in self-regulated learning. *Handbook of self-regulation/Academic*.

Pintrich, P. R., Wolters, C. A., & Baxter, G. P. (2000). 2. Assessing Metacognition and Self-Regulated Learning.

Prevatt, F., Proctor, B., Best, L., Baker, L., Van Walker, J., & Taylor, N. W. (2012). The positive illusory bias: Does it explain self-evaluations in college students with ADHD? *Journal of Attention Disorders, 16*(3), 235-243.

Prinz-Weiß, A., Lukosiute, L., Meyer, M., & Riedel, J. (2023). The role of achievement emotions for text comprehension and metacomprehension. *Metacognition and Learning, 18*(2), 347-373.

Rawson, K. A., Dunlosky, J., & Thiede, K. W. (2000). The rereading effect: Metacomprehension accuracy improves across reading trials. *Memory & cognition, 28,* 1004-1010.

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

Reinnard, T., (2021). Molekularbiologische Methoden 2.0. [Methods in molecular biology] UTB.

Repovš, G., & Baddeley, A. (2006). The multi-component model of working memory: Explorations in experimental cognitive psychology. *Neuroscience, 139*(1), 5-21.

Rhodes, M. G., & Tauber, S. K. (2011). The influence of delaying judgments of learning on metacognitive accuracy: a meta-analytic review. *Psychological Bulletin, 137*(1), 131.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin, 86*(3), 638.

Rowland, C. A. (2014). The effect of testing versus restudy on retention: a meta-analytic review of the testing effect. *Psychological Bulletin, 140*(6), 1432.

Sabel, J. L., Dauer, J. T., & Forbes, C. T. (2017). Introductory biology students' use of enhanced answer keys and reflection questions to engage in metacognition and enhance understanding. *CBE—Life Sciences Education, 16*(3), ar40.

Schlottke, P. F., Strehl, U., & Christiansen, H. (2019). Aufmerksamkeitsstörung. *Lehrbuch der Verhaltenstherapie, Band 3: Psychologische Therapie bei Indikationen im Kindes- und Jugendalter. [Textbook of Behavioural Therapy. Psychological Therapy for children and youth]*, 429-451.

Schneider, W. (1986). The role of conceptual knowledge and metamemory in the development of organizational processes in memory. *Journal of Experimental Child Psychology, 42*(2), 218-236.

Schraw, G. (1998). Promoting general metacognitive awareness. *Instructional Science, 26*(1), 113-125.

Schroeder, N. L., Nesbit, J. C., Anguiano, C. J., & Adesope, O. O. (2018). Studying and constructing concept maps: A meta-analysis. *Educational Psychology Review*, *30*, 431-455.

Schuster, C., Stebner, F., Leutner, D., & Wirth, J. (2020). Transfer of metacognitive skills in self-regulated learning: an experimental training study. *Metacognition and Learning*, *15*(3), 455-477.

Schwartz, B. L. (2024). Inferential theories of retrospective confidence. *Metacognition and Learning*, 1-32.

Sebesta, A. J., & Bray Speth, E. (2017). How should I study for the exam? Self-regulated learning strategies and achievement in introductory biology. *CBE—Life Sciences Education*, *16*(2), ar30.

Serra, M. J., & Dunlosky, J. (2010). Metacomprehension judgements reflect the belief that diagrams improve learning from text. *Memory*, *18*(7), 698-711.

Sonuga-Barke, E. J. (2003). The dual pathway model of AD/HD: an elaboration of neuro-developmental characteristics. *Neuroscience & Biobehavioral Reviews*, *27*(7), 593-604.

Spruce, R., & Bol, L. (2015). Teacher beliefs, knowledge, and practice of self-regulated learning. *Metacognition and Learning*, *10*, 245-277.

Stanton, J. D., Dye, K. M., & Johnson, M. S. (2019). Knowledge of learning makes a difference: A comparison of metacognition in introductory and senior-level biology students. *CBE—Life Sciences Education*, *18*(2), ar24.

Stanton, J. D., Halmo, S. M., Carter, R. J., Yamini, K. A., & Ososanya, D. (2024). Opportunities for guiding development: insights from first-year life science majors' use of metacognition. *Journal of Microbiology and Biology Education*, e00053-00024.

Stanton, J. D., Neider, X. N., Gallegos, I. J., & Clark, N. C. (2015). Differences in metacognitive regulation in introductory biology students: when prompts are not enough. *CBE—Life Sciences Education*, *14*(2), ar15.

Stanton, J. D., Sebesta, A. J., & Dunlosky, J. (2021). Fostering metacognition to support student learning and performance. *CBE—Life Sciences Education*, *20*(2), fe3.

Stroebe, W., Postmes, T., & Spears, R. (2012). Scientific misconduct and the myth of self-correction in science. *Perspectives on Psychological Science*, *7*(6), 670-688.

Tanner, K. D. (2012). Promoting student metacognition. *CBE—Life Sciences Education*, *11*(2), 113-120.

Taranto, D., & Buchanan, M. T. (2020). Sustaining lifelong learning: A self-regulated learning (SRL) approach. *Discourse and Communication for Sustainable Education*, *11*(1), 5-15.

Thiede, K. W., Anderson, M., & Therriault, D. (2003). Accuracy of metacognitive monitoring affects learning of texts. *Journal of Educational Psychology*, *95*(1), 66.

Thiede, K. W., Dunlosky, J., Griffin, T. D., & Wiley, J. (2005). Understanding the delayed-keyword effect on metacomprehension accuracy. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(6), 1267.

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.

UNESCO Institute for Statistics, U. I. f. (2003). *PISA Literacy Skills for the World of Tomorrow Further Results from PISA 2000: Further Results from PISA 2000*. OECD Publishing.

Veenman, M. V., Van Hout-Wolters, B. H., & Afflerbach, P. (2006). Metacognition and learning: Conceptual and methodological considerations. *Metacognition and Learning*, *1*, 3-14.

Verhoeff, R. P., Knippels, M.-C. P., Gilissen, M. G., & Boersma, K. T. (2018). The theoretical nature of systems thinking. Perspectives on systems thinking in biology education. In *Frontiers in Education* (Vol. 3, p. 40). Frontiers Media SA.

Volz-Sidiropoulou, E., Boecker, M., & Gauggel, S. (2016). The positive illusory bias in children and adolescents with ADHD: further evidence. *Journal of Attention Disorders*, *20*(2), 178-186.

Von Wright, G.H. (1963). *The varieties of goodness.* Routlege and Kegan Paul.

Willcutt, E. G. (2012). The prevalence of DSM-IV attention-deficit/hyperactivity disorder: a meta-analytic review. *Neurotherapeutics*, *9*(3), 490-499.

Winne, P. H. (2011). A cognitive and metacognitive analysis of self-regulated learning: Faculty of education, Simon Fraser University, Burnaby, Canada. In *Handbook of self-regulation of learning and performance* (pp. 29-46). Routledge.

Zechmeister, E. B., & Shaughnessy, J. J. (1980). When you know that you know and when you think that you know but you don't. *Bulletin of the Psychonomic Society*, *15*(1), 41-44.

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. (1989). A social cognitive view of self-regulated academic learning. *Journal of Educational Psychology*, *81*(3), 329.

Zimmerman, B. J. (1990). Self-regulated learning and academic achievement: An overview. *Educational Psychologist*, *25*(1), 3-17.

Zimmerman, B. J., & Moylan, A. R. (2009). Self-regulation: Where metacognition and motivation intersect. In *Handbook of Metacognition in Education* (pp. 299-315). 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), 121-1

Appendix

List of Figures

Study I

Study II

Study III

List of Tables

Study I

Study II

Study III