

# Using digital formative assessments to improve learning in physics education

Citation for published version (APA):

Molin, F. M. B. M. (2022). *Using digital formative assessments to improve learning in physics education*. [Doctoral Thesis, Maastricht University]. ROA. <https://doi.org/10.26481/dis.20220114fm>

## Document status and date:

Published: 01/01/2022

## DOI:

[10.26481/dis.20220114fm](https://doi.org/10.26481/dis.20220114fm)

## Document Version:

Publisher's PDF, also known as Version of record

## Please check the document version of this publication:

- A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
- The final author version and the galley proof are versions of the publication after peer review.
- The final published version features the final layout of the paper including the volume, issue and page numbers.

[Link to publication](#)

## General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license above, please follow below link for the End User Agreement:

[www.umlib.nl/taverne-license](http://www.umlib.nl/taverne-license)

## Take down policy

If you believe that this document breaches copyright please contact us at:

[repository@maastrichtuniversity.nl](mailto:repository@maastrichtuniversity.nl)

providing details and we will investigate your claim.

Using Digital Formative Assessments  
to Improve Learning in Physics Education

© François Molin, 2022

All rights reserved. No part of this publication may be reproduced, stored in an automated data system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without prior permission of the author.

Published by ROA  
Postbus 616  
6200 MD Maastricht

ISBN: 978-90-5321-605-7

Printed in the Netherlands by Canon

**Using Digital Formative Assessments  
to Improve Learning in Physics Education**

**DISSERTATION**

to obtain the degree of Doctor  
at Maastricht University,  
on the authority of Rector Magnificus,  
Prof. dr. R.M. Letschert  
in accordance with the decision of the Board of Deans,  
to be defended in public on  
Friday 14 January 2022, at 10:00 hours

by

**François Maurits Bertine Maria Molin**

**Promotors**

Prof. dr. C.M.G. Haelermans

Prof. dr. W.N.J. Groot

**Co-promotor**

Dr. S.J.J. Cabus (KU Leuven)

**Doctoral thesis committee**

Prof. dr. T. Schils (Chair)

Prof. dr. W.F. Admiraal (Leiden University)

Prof. dr. P.H.M. Drijvers (Utrecht University)

Prof. dr. J.J.G. van Merriënboer

Prof. dr. I.F. de Wolf

## Acknowledgements

There are several people who have helped me in the realization of this dissertation over the past four years. A number of them I would like to thank in particular. First of all, many thanks to my two supervisors Prof. dr. Carla Haelermans and Prof. dr. Wim Groot and my co-supervisor Dr. Sofie Cabus, who made it possible for me to successfully complete this process. Your excellent supervision has made sure that we have been able to make some valuable scientific contributions during my PhD research. Carla, thank you very much for everything you have done for me. Your guidance has in fact played a major role in the realization of this dissertation. Thank you for all your help, consideration and interesting conversations we had when we thought about the research together. Your feedback has helped me to bring the logic and coherence to a higher level, while your enthusiasm and expertise gave me the confidence to investigate new topics in my research. Wim, thank you for the faith you have in me and for giving me the opportunity to work as an external PhD candidate at TIER. Our meetings and the feedback were important to me; these were moments in which there was both appreciation for my efforts and also opportunities for further development. I really valued the cooperation with two supervisors. Sofie, you sparked my enthusiasm for scientific research and stimulated me (perhaps unintentionally) to do PhD research. For me, you were the methodological conscience who knew how to distil the imperfections in the analyses from my work. You always thought actively and sharply in solutions. Your approach towards doing research has made me more critical. Thank you for that! Carla, Sofie and Wim, I learned a lot and am grateful that you were willing to be my supervisors and co-supervisor. I look forward to the opportunities to continue to work together in educational research.

Of course, I would also like to thank my employer Onderwijsgemeenschap Venlo & Omstreken (OGVO) and the school management of College Den Hulster for offering me a PhD task. This was a unique opportunity to develop myself and I still find it special that my employer made this possible for me.

I would also like to thank the members of the reading committee, Prof. dr. Trudie Schils, Prof. dr. Inge de Wolf, Prof. dr. Wilfried Admiraal, Prof. dr. Paul Drijvers and Prof. dr. Jeroen

## Acknowledgements

van Merriënboer for taking time to read my dissertation and for their valuable comments and questions.

Furthermore, I would like to thank everyone who contributed to the realization of this dissertation. First of all, my thanks goes to Prof. dr. Anique de Bruin for the inspiring discussions and her contribution to the study described in Chapter 5. I am sure this chapter will soon be published in a matching journal. In addition, the research in this dissertation would never have been possible without the participation of many colleagues in secondary education. My appreciation and thanks therefore go to all those colleagues who participated in the studies: Herman Franssen, Louis Lenders and Han Jeurissen of College Den Hulster, Monique Wittebrood, Hans Dreyer and Erik van de Leur of Blariacum College, Rick Cremers, Dennis Basten, Rob Killaars and Jordi Janssen of Valuas College, Joyce Dolfsma and Omar Chouhabi of Bouwens van der Boijecollege, and Natascha Musters of Norbertus Gertrudis Lyceum. My thanks also go to all the students from these schools who participated in the studies.

Furthermore, within TIER, where it all started, my thanks go to all colleagues for the good time. Astrid Lamers, Carla Lenssen, Dr. Joris Ghysels, Prof. dr. Kristof de Witte and all fellow PhD candidates with whom I shared "promotion joys and sorrows". I would especially like to thank my two PhD roommates and paranympths Mélanie Monfrance and Melline Somers for their always sincere interests and helpfulness. You two are really great!

There is no space here to mention each one by name, but the colleagues and students of College Den Hulster have each in their own way inspired and supported me in the research. Thank you for that! In particular, I would like to thank my colleague Kevin Bamforth for reading through my papers and dissertation. I could count on you and nothing was too much for you. You are a wonderful colleague! I also thank Jules Storcken, René Dohmen, Ralph van Bergen, Har Peters and everyone who sometimes asked me what exactly I was doing in this research for their kind interest.

In conclusion, I would like to thank my parents and my family; my wife Desirée and my children Gilles and France. Thank you for the support you have given me and the endless patience you have had over the past four years, because I wanted to do research so much. Your support made this great time even better than it already was. Thank you for this!







# Contents

Acknowledgements.....	v
List of Tables.....	xi
List of Figures.....	xii
<b>Chapter 1 General Introduction .....</b>	<b>1</b>
1.1 Introduction.....	2
1.2 Background.....	5
1.3 Aims and contributions of this dissertation .....	12
1.4 Dissertation structure and outline .....	14
<b>Chapter 2 Toward Reducing Anxiety and Increasing Performance in Physics Education .....</b>	<b>19</b>
2.1 Introduction.....	21
2.2 Theoretical background .....	22
2.3 Methods .....	25
2.4 Results .....	33
2.5 Conclusion and discussion.....	38
<b>Chapter 3 Do Feedback Strategies Improve Students' Learning Gains? .....</b>	<b>43</b>
3.1 Introduction.....	45
3.2 Literature.....	47
3.3 Materials and methods .....	51
3.4 Results .....	63
3.5 Conclusion and discussion.....	71
Appendix 3.1.....	74
<b>Chapter 4 The Effect of Feedback on Metacognition .....</b>	<b>77</b>
4.1 Introduction.....	79
4.2 Conceptual framework and literature.....	82
4.3 Materials and methods .....	87
4.4 Results .....	98
4.5 Conclusion and discussion.....	109
Appendix 4.1.....	113
Appendix 4.2.....	114
Appendix 4.3.....	115
<b>Chapter 5 A Conceptual Framework to Understand Learning: The Role of Prompts and Diagnostic Cues .....</b>	<b>117</b>
5.1 Introduction.....	119
5.2 Theory on metacognition: monitoring and control .....	120

## Contents

5.3	Diagnostic cues to monitor learning .....	122
5.4	A conceptual framework .....	124
5.5	Diagnostic cues in formative assessments with SRS .....	125
5.6	Prompts in formative assessments with SRS .....	130
5.7	Utilization problems .....	136
5.8	Conclusion .....	137
<b>Chapter 6 Conclusion and Discussion .....</b>		<b>141</b>
6.1	Conclusion and discussion.....	142
6.2	Study limitations and future research.....	151
6.3	Practical implications .....	153
6.4	Recommendations .....	154
<b>Chapter 7 Impact Statement .....</b>		<b>157</b>
References.....		163
Summary .....		189
Samenvatting.....		197
About the Author .....		205
ROA Dissertation Series.....		207

## List of Tables

Table 2.1: Mean differences on standardized pre-treatment characteristics .....	33
Table 2.2: Multilevel regression analyses predicting post-score anxiety and academic performance	35
Table 3.1: T-tests between students in the cooperative condition .....	57
Table 3.2: Descriptive statistics of answered question pairs .....	57
Table 3.3: Comparison between untreated and treated conditions .....	58
Table 3.4: Multivariate regression analyses predicting learning gains .....	64
Table 3.5: Robustness analyses with outcome learning gain .....	66
Table 3.6: Multilevel regression analyses predicting learning gains .....	75
Table 4.1: Descriptive statistics of the sample .....	96
Table 4.2: ANOVA results for pre-tests between conditions .....	96
Table 4.3: Regression analyses predicting standardized post-score metacognition .....	101
Table 4.4: Regression analyses predicting standardized post-score motivation .....	105
Table 4.5: Regression analyses predicting mediation analysis metacognition .....	108
Table 4.6: T-tests between included and excluded students .....	114
Table 4.7: Regression analysis post-score metacognition .....	115
Table 6.1: Main findings of all research questions .....	144

## List of Figures

Figure 1.1: Contents overview .....	16
Figure 2.1: Course and design of the intervention .....	26
Figure 2.2: Research design .....	32
Figure 2.3: Mediation analysis .....	38
Figure 3.1: Description of the experimental design .....	53
Figure 3.2: Two examples of two paired questions used in this study .....	55
Figure 3.3: Flowcharts of untreated and treated conditions .....	68
Figure 3.4: Flowcharts of two subgroups in the cooperative condition .....	70
Figure 4.1: Model of relations .....	82
Figure 4.2: Experimental design .....	90
Figure 4.3: Overview of the timeline of the experiment .....	92
Figure 4.4: Mean standardized post-scores of metacognition .....	100
Figure 4.5: Mean standardized post-scores of motivation .....	104
Figure 4.6: Mediation analysis of the cooperative and individual treatment .....	107
Figure 4.7: Two screenshots of a paired question in <i>Socratic</i> on smartphones screens .....	113
Figure 5.1: Relation between monitoring and control .....	121
Figure 5.2: A developed framework of prompts and cues by formative assessments with SRS .....	127





Chapter 1  
General Introduction



## 1.1 Introduction

Science plays an important role in our social and economic lives. Many situations in everyday life require some level of understanding of science-related reasoning or science-related tools before they can be fully comprehended and managed. With this central role of science, today's society demands that all citizens, and not only those who work in science, are to some extent scientifically, mathematically and technologically educated. Understanding theories of basic scientific principles and the skills to solve basic scientific problems are more important than ever (Chang & Chiu, 2005; OECD, 2007).

Scientific innovations and technological changes are fundamental sources of economic growth for countries (OECD, 2007; Williams, 2003). A country's role in tomorrow's international competition depends to a large extent on students' performance in science-related subjects in schools (Dolin & Krogh, 2010; OECD, 2007). Well-performing science students with a sufficiently high level of scientific knowledge maintain the supply of scientists and technically skilled employees, which contributes to a country's economic prosperity (Laugksch, 2000). While governments focus on stimulating students in science-related fields, there are concerns about a decline in enrolment in science-related studies and a shortage of science graduates on the labour market (OECD, 2016). This is not without reason. Despite the fact that most people recognize the important role of science in daily life, a large proportion of students experiences science as inaccessible. They feel uncomfortable with it and have little confidence in their own ability to do science. Science is strongly stereotyped as too difficult and unattractive, creating the image that only the best students with the highest grades can succeed (Langen & Dekkers, 2005; Marnell, 2012; Seymour & Hewitt, 1997). It is an abstract and difficult discipline which, as it is argued, is accessible only to a limited group of students who are capable of conducting abstract and complex reasoning processes. This view is widely accepted, particularly in the case of physics, which is often considered as the most abstract and difficult of all sciences (White & Frederiksen, 1998). A large number of students in upper secondary education avoid STEM-related profiles or trajectories, or do not continue a science related study after secondary education as they have doubts about their science abilities (Hong, 2009; Zemira & Bracha, 2014). A major reason for avoiding science-related studies is the way of thinking of students; about half of the population of students are concrete thinkers,

while science requires formal thinking (Barnes, 1977; Cohen, Hillman & Agne, 1978; Udo, Ramsey & Mallow, 2004). Being unable to think formally may lead to science anxiety and avoidance of science enrolment. The small number of students in some science-related studies, such as physics, is an unfortunate consequence (Udo et al., 2004).

Secondary education plays an important role in students' choice of studies related to sciences. To boost the numbers enrolled in science-related studies, education must give students the confidence that they are able to understand science and to reason about it with the 'right' level of thinking. For this purpose, teachers must motivate students, as motivation in science strongly relates to students' aspirations for future science-related studies and careers (Hong, 2009; OECD, 2007). To achieve the goal of more enrolled students, it is argued that teachers should also focus on providing frequent, meaningful feedback in classes (Champagne & Newell, 1992). Meaningful feedback has a powerful influence on students' understanding and way of thinking (Black & William, 1998a; Hattie & Timperley, 2007). It motivates students to learn and helps them in the short term to generate new knowledge, and in the long term to improve metacognitive awareness and learning performance (Egelandsdal & Krumsvik, 2019a; Jones, 2007; Jonsson, 2013). Metacognitively aware students know what to do when they do not know what to do; that is, they have strategies for finding out or figuring out what they need to do (Anderson, 2002, p. 2). Students gain more confidence in their own abilities, as meaningful feedback leads to more formal and deeper thinking (Chin, 2006; Erdogan & Campbell, 2008; Voerman, Meijer, Korthagen & Simons, 2012). In addition, using meaningful feedback to show students why an answer is right or wrong may reduce their anxiety and encourages them to achieve a higher grade in a next exam (Arkin & Schumann, 1984; Rocklin & Thompson, 1985; Sullivan, 2017).

However, despite its importance, students in traditional classrooms generally hardly receive meaningful feedback about their own understanding of course content. It is found that this is due to the fact that teachers (1) do not have enough time to provide feedback due to overloaded programmes or overcrowded classrooms; (2) do not know which students do or do not understand the course content; or (3) do not have the knowledge and skills to assess students' understanding of content without grading (Dudaité & Prakapas, 2017; Ketabi & Ketabi, 2014; Lee, Irving, Pape & Owens, 2015; Trees & Jackson, 2007).

To give students more insight and confidence in their own scientific abilities, it is important that teachers gain insight into students' learning needs through ongoing formative assessments in classrooms (Levesque, 2011; Lopez, Love & Watters, 2014; Núñez-Peña, Bono & Suárez-Pellicioni, 2015). These formative assessments consist of weekly, interactive checks of students' understanding, providing meaningful feedback, and adapting teaching strategies to meet students' needs (Nicol & Macfarlane-Dick, 2006). Information and communication technology (ICT) that collects accurate real-time data plays a major role in this respect (Mostafa, Echazarra & Guillou, 2018; Zemira & Bracha, 2014). Using ICT helps teachers to conduct easy-to-organize formative assessments with teacher-student and peer interactions (Narciss, 2008; Wong & Yang, 2017). ICT-supported formative assessments actively encourage students to answer questions with their devices, participate in peer discussions, ask questions to the teacher and reflect on their own learning, while teachers build on student ideas, provide meaningful feedback to move students forward in their learning and make instructional decisions about subsequent lessons (Furtak et al., 2016). Here, the ultimate goal is that students develop their own 'learning to learn' skills, also referred to as metacognitive skills (OECD, 2005). In developing these skills, students experience which learning behaviors are most effective for them, learn to assess themselves and see which learning strategies are most effective. It is for these reasons that formative assessments are perhaps one of the most effective interventions in education for enhancing academic performance (OECD, 2005). Black and Wiliam (1998a) described the gains in achievement in formative assessments as "*the largest ever reported for educational interventions*" (p. 61).

To conclude, most of the literature that studied ICT-supported formative assessments has focused on the extent to which feedback affects academic performance (Chien, Chang & Chang, 2016). As the way in which teachers organize assessments affects students' behavior and thought processes, it is important to gain insight into how feedback affects learning in everyday classrooms.

## 1.2 Background

Meaningful feedback is one of the most powerful means to increase student learning (Hattie & Gan, 2011; Hattie & Timperley, 2007). Many researchers have demonstrated that feedback has positive effects on learning outcomes. Although there is no widely approved model of how feedback increases learning outcomes, most research has demonstrated that students confirm or modify their understanding and skills after receiving feedback on their answers. It is considered as information about how successful something has been or is being done. Hattie and Timperley (2007) defined feedback as *“information provided by an agent (e.g. teacher, peer, book, parent, self, experience) regarding aspects of one’s performance or understanding”* (p. 81). In a traditional classroom setting, feedback is given by teachers who know what knowledge and skills need to be studied or trained, recognize good performances, and show how poor performances can be improved into good performances. Here, the timing of feedback plays an important role. Regarding this timing, a distinction can be made between immediate and delayed feedback. Immediate feedback, as Shute (2008) stated, is delivered right after a student answers a question or completes an item. Delayed feedback is given a few hours, a day, or even a week after a student answers a question. Previous literature has suggested that students prefer immediate over delayed feedback, as students perceive immediate feedback to be most useful for learning (Miller, 2009). They tend to ignore feedback when it is not provided in a timely way. So feedback could be effective (1) when it contains suggestions to improve performance, and (2) when it is timely or immediate.

Most feedback moments in a classroom occur spontaneously in an unplanned way. Despite the value of these moments to student learning, there is evidence that students prefer planned classroom activities in which feedback is provided (e.g. Iverson, Iverson & Lukin, 1994). For this purpose, teachers can choose to integrate frequent formative assessments into their teaching as a vehicle for providing immediate feedback in a planned way (Black & Wiliam, 1998a; Hattie & Timperley, 2007; Shute, 2008). Formative assessments are interactive assessments that evaluate students’ progress and understanding on a frequent basis. Formative assessments are defined by Black and Wiliam (1998a) as *“activities undertaken by teachers, and/or by their students, which provide information to be used as feedback to modify the teaching and learning activities in which they are engaged”* (p. 7-8). This means that not

only students receive feedback on their understanding to make decisions in their learning, but also teachers receive feedback on students' understanding to adjust their teaching to the perceived learning needs. Many studies argue that formative assessments with immediate feedback have positive effects on student outcomes (e.g. Black & Wiliam, 1998a; Black & Wiliam, 1998b). The effect size in 23 meta-analyses is 0.73 and 0.90 for feedback and formative assessments, respectively (Hattie & Learning, 2009). The effect size for formative assessments is the third highest of the 138 achievement influencing factors.

Formative assessment is a process of collecting data on students' understanding, skills, and learning progress. One way to collect data and support the assessment process is by using ICT, like mobile technological tools (Shirley & Irving, 2015; Wong & Yang, 2017). Technological tools in educational settings are often referred as 'polling technologies' or 'student response systems'. For the sake using a consistent term, in this dissertation we use the term *student response systems* (SRS) when referring to all categories of technological tools. These SRS are communication mediums that provide teachers and students with quick information as students submit responses to the system. This system provides students with information that allows them to compare their personal responses with the collective responses displayed in front of the class, while allowing teachers to adapt their instruction to the needs of their students. Technology supported formative assessments have the advantage of (1) delivering feedback to students while taking an assessment; (2) providing immediate information of students' understanding; and (3) providing a high assessment efficiency by offering multiple questions in a limited amount of time, without intentionally increasing the workload for teachers. It is argued by Hattie and Timperley (2007) that the first advantage has a positive effect on students' learning outcomes<sup>1</sup>, because 'the gap between the current and desired understanding' can immediately be reduced. The graphical representation of answers in front of the class make feedback accessible for teachers. The role of the teacher is now crucial, as

---

<sup>1</sup> Since Mazur (1997) introduced SRS into his physics courses, many studies have been conducted that examined the effects of formative assessments with SRS on academic performance. Most of these studies show variations in performance outcomes. For example, the studies of Mayer et al. (2009) and Hubbard and Couch (2018) both showed large improvements, while Morling, McAuliffe, Cohen and DiLorenzo (2008) and Fortner-Wood, Armistead, Marchand and Morris (2013) showed small improvements and Patterson, Kilpatrick and Woebkenberg (2010) and Sutherlin, Sutherlin and Akpanudo (2013) showed no effects. These variations will not be due to chance, but suggest that academic performance depends on how an assessment is organized.

teachers encourage students to improve the clarity and quality of their explanations, eliminate incorrect ways of thinking, and help students to be aware of what they should focus their learning on (Gipps, 2005). For these reasons, teachers' feedback has been positively related to student learning (Hattie & Timperley, 2007; Kluger & DeNisi, 1996).

As to the way teachers organize an assessment, they can also choose a more active form of formative assessment with more interactions between students. This so called 'peer discussion' approach is an evidence-based instructional strategy, based on the work of Mazur (1997). This strategy is well-known and widely used; especially in science classrooms, where it is noted as perhaps the most important element of the assessment process (Caldwell, 2007; Crouch & Mazur, 2001; Knight & Brame, 2018; Tullis & Goldstone, 2020). In peer discussions, teachers pose questions to students and students discuss their responses with their peers. It creates a dynamic learning environment in which students have the opportunity to be coached by peers. The peer discussions help students to connect their ideas to others' ideas, where students learn from each other (Crouch & Mazur, 2001; Crouch, Watkins, Fagen & Mazur, 2007; Lasry, Mazur & Watkins, 2008).<sup>2</sup>

The meta-analysis of Bangert-Drowns, Kulik, Kulik and Morgan (1991) demonstrated that frequent formative assessments increase student learning. They suggested that the benefits of frequent formative assessments, in addition to feedback, depend on several variables. The variables themselves are not mentioned by Bangert-Drowns et al. (1991), but previous literature has suggested that frequent formative assessments reduce anxiety (Batchelor, 2015; Fulkerson & Martin, 1981), increase motivation (Black & Wiliam, 1998a; Doucet, Vrins & Harvey, 2009) and increase metacognitive awareness (Brady, Seli & Rosenthal, 2013; Jones, Antonenko & Greenwood, 2012). To better understand the impact of frequent formative

---

<sup>2</sup> Formative assessments with SRS occur primarily in classrooms, with personal contact between teacher and students. However, the COVID-19 pandemic forced an abrupt switch to online formative assessments, with teacher and students separated by space and/or time. As a result, the interaction and communication between teachers and students or between students themselves, immediate feedback, and respect for differences between students' skills and learning styles have declined (Khan & Jawaid, 2020; Tartavulea, Albu, Albu, Dieaconescu & Petre, 2020). Teachers indicate that there is less feedback in these online formative assessments, which is a result of reduced interaction due to fewer or missing facial expressions and gestures, and fewer questions and reflections from students (Willermark, 2021). This demonstrates the importance of a dynamic learning environment, with face-to-face contact and interactions between teacher and students, and students among themselves.

assessments in classrooms on student learning, the aspects of anxiety (and in this dissertation science anxiety), motivation and metacognition need to be taken into account and further examined.

### **1.2.1 Science anxiety**

Science anxiety is an individual emotional state that impedes the learning of sciences and leads to negative science-related attitudes and self-perceptions. It occurs when a feeling of worry and tension is experienced when faced with science. Science anxiety is an important factor that negatively affects science performance, as science-anxious students are unable to concentrate on science exams, because they feel insecure about their own abilities (Cassady & Gridley, 2005). Science-anxious students have mostly poor science skills, utilize ineffective study strategies in an exam preparation, worry over potential failures and become frustrated, which all negatively affect understanding of science content (Cassady, 2004; Naveh-Benjamin, McKeachie, Lin & Holinger, 1981).

#### *1.2.1.1 Reducing (science) anxiety through feedback and formative assessments*

There is evidence that providing item feedback in formative assessments decreases anxiety. Providing immediate meaningful feedback after answering an item lowers levels of worry, tension and anxiety (Fulmer, 1976; Rocklin & Thompson, 1985). The study by Rocklin and Thompson (1985) showed that item feedback improves students' performance, which is a result of a reduction in anxiety. To reduce the impact of science anxiety on students' performance, teachers can offer non-threatening (non-graded) formative assessments. Formative assessments have the advantage that students can take an assessment in a less stressful environment, since answering questions correctly or not will not influence course grades. It is argued that students experience less anxiety in formative assessments than in summative assessments (Cassady, Budenz-Anders, Pavlechko & Mock, 2001), as students can identify their strengths and increase their self-confidence without feeling the pressure of an appraisal. The formative assessments act as exam preparation activities, where students perform tasks that are comparable to tasks in subsequent summative assessments.

The use of technological-aided assessments has benefits for (anxious) students, as the student response systems (SRS) allow teachers and peers to provide immediate feedback after each

answered item. Another reason is that the SRS allow students to submit anonymous responses to the system. This means that students actively participate in answering a question, but that their actual responses are not revealed to other students, and are not immediately obvious to the teacher. This anonymity creates a safe learning environment for students without peer or teacher pressure and releases them from nervousness and anxiety, resulting in more active participation of students (Brady et al., 2013; Yu, Chen, Kong, Sun & Zheng, 2014).

### **1.2.2 Motivation**

Active participation is also a result of motivation. The Latin derivative of motivation means “*to move*”. Motivation is an important factor that stimulates students’ learning by focusing attention on the task, which results in an improvement of learning outcomes (Pekrun, 2006; Pintrich & de Groot, 1990). Focusing attention on a task is a consequence of an intrinsic motivation to perform that task (Eccles, 2005). A lack of focusing attention implies that intrinsic motivation, or the motivation to work on a task primarily for its own benefit, is lacking and that extrinsic motivation, or the motivation to engage in a task because it is a means to an end, is necessary to sustain the focus of attention on the task (Deci, Ryan & Williams, 1996; Pekrun, Goetz, Daniels, Stupnisky & Perry, 2010; Sansone & Thoman, 2005).

#### *1.2.2.1 Increasing motivation through feedback and formative assessments*

Providing meaningful feedback affects student motivation in a positive way (Black & Wiliam, 1998a; Shute, 2008), especially when feedback provides students with specific information about their current state of understanding with respect to what is expected, or when feedback informs students of how to improve their performance on tasks (Dresel & Haugwitz, 2008; Moreno, 2004; Shute, 2008). In daily teaching practice, the effect of feedback on motivation depends not only on what is taught to students, but also how it is taught (Koka & Hein, 2006; Pat-El, Tillema & van Koppen, 2012). Feedback motivates students when a teacher considers students’ perspectives, acknowledges their feelings, and gives them meaningful information to continue learning (Pat-El et al., 2012). If so, creating ‘sufficient’ (e.g. weekly) opportunities to provide meaningful feedback is essential to keep students motivated. Too few or insufficient opportunities diminishes students’ motivation to learn (Black & Wiliam, 1998a; Kluger & DeNisi, 1996). Teachers can choose to integrate formative assessments into their teaching as a means to provide frequent and sufficient feedback. Formative assessments are



considered as one of the most effective tools to improve student motivation in classrooms (Cauley & McMillan, 2010). These assessments help students to focus attention on tasks and increase interest to perform tasks, resulting in an increased motivation and satisfaction of students' basic needs to feel competent (Chien et al., 2016; Ryan & Deci, 2000). From this point of view, there is substantial evidence demonstrating a positive relationship between formative assessment and students' motivation and performance (Black & Wiliam, 1998). One way to boost students' motivation in formative assessments is by using SRS. These systems create opportunities for teachers by making the daily teaching practices more attractive for students, motivating them to take part in assessments, and increasing classroom interactions and a positive sense of community (Buil, Catalán & Martínez, 2016; Caldwell, 2007; Kay & LeSage, 2009). It is the integration of SRS into lessons and assessments that today's students expect in order to stay interested, focused and motivated (Smart, Kelley & Conant, 1999; Williamson, Sprague & Dahl, 2010).

### **1.2.3 Metacognition**

Metacognition refers to students' ability to monitor and control cognitive processes. It is typified as one's cognition about cognition, or one's awareness and organization of cognitive skills. The term metacognition has many definitions depending on how it is used (e.g. Pintrich, 2000a; Winne & Hadwin, 2008; Zimmerman, 2000), but an important commonality is that metacognition always refers to monitoring and controlling cognitive processes. The distinction between cognitive processes and metacognitive processes is not always easy to make. The prefix 'meta' is added to indicate that metacognition goes beyond cognition. Cognitive processes are used when performing a task, while metacognitive processes are used to understand how the task was performed. So, answering a question is a cognitive process, but reflecting on the correctness of an answer and realizing that the solution strategy has been applied correctly or incorrectly is a metacognitive process.

Metacognition plays an important role in affecting student learning (Dunlosky & Thiede, 1998; Dunning, Johnson, Ehrlinger & Kruger, 2003) and is found to be critical in learning sciences (Mota, Körhasan, Miller & Mazur, 2019; Taasobshirazi, Bailey & Farley, 2015). A reason for this is that physics, biology and chemistry are relatively unfamiliar sciences for secondary students, so students need to be proactive, curious, and self-regulated during the process of

understanding. Another reason is that science courses require metacognitive skills that enable students to identify, define, mentally imagine, and plan how to tackle problems before using equations to solve them quantitatively. The studies of Neto and Valente (1997) and Rozencwajg (2003) both demonstrated that students who show a variety of metacognitive skills while solving science problems (e.g. physics problems) are more likely to correctly solve science problems and problems in general.

It is important for students to increase their metacognitive awareness. Students who receive instruction in how to solve problems and questions develop metacognitive skills that make them more successful in their professional careers. They gain understanding of how they learn, how they process information, and how they memorize, making them better able to create situations for themselves that facilitate learning.

#### *1.2.3.1 Improving metacognition through feedback and formative assessments*

Metacognitive skills can be improved in classrooms by creating a learning environment in which students demonstrate, explain, discuss, and control their own thought processes. To do this, teachers need to engage students in metacognitive activities, such as tasks that stimulate them to think about how to address problems while completing these tasks. Formative assessments are metacognitive activities that help students develop a range of effective learning strategies and skills that are invaluable for lifelong learning (OECD, 2005). The feedback provided is an essential part of the learning process as it encourages students to reflect on their own thinking. It identifies students' strengths and weaknesses of problem approaching and provides them with information about how they handled the task and how they can constructively change their approach to the next problem. The use of SRS can support the metacognitive nature of formative assessments by providing students with *immediate* feedback. The instantaneous graphical feedback from ICT facilitates students' ability to monitor themselves relative to other students in class. This encourages students' active involvement in the learning process by reinforcing their metacognitive skills to monitor their understanding and that of peers (Lee, Irving, Pape & Owens, 2015).

All in all, the previous literature has shown that formative assessments with SRS have the potential to reduce students' science anxiety, increase students' motivation and improve

students' metacognitive awareness. The next section will describe how these three variables in this dissertation contribute to the already existing literature.

### 1.3 Aims and contributions of this dissertation

In daily science education, the majority of teachers use formative assessments without realizing that the way an assessment is organized and the way feedback is provided affect student learning. A result is that most of the interactions in a formative assessment are completely controlled by the teachers themselves (Jurik, Gröschner & Seidel, 2013). There is limited room for conversations and teacher feedback is mostly a monologue with a binary, *"get it or don't"*, perspective (Furtak et al., 2016). On top of that, most science teachers are also not aware that the way an assessment is organized and the way feedback is provided affect science anxiety, motivation to learn and metacognitive awareness; three variables that González, Fernández and Paoloni (2017) argue are driving forces of learning in science courses. For these reasons, the aim of this dissertation is to provide teachers with (1) evidence that the way in which feedback is provided affects science anxiety, motivation, metacognitive awareness, and students' performance, and (2) insight in that how an assessment is organized affects the number of prompts (activities such as feedback) and diagnostic cues (information that students use) that finally result in an enhancement of student learning.

Although previous studies have provided important insights into student learning, there is unfortunately still a lack of evidence as to why performance improvements occur when students are formatively assessed with SRS. The theoretical background described in previous sections shows that feedback is an important aspect of academic performance, but that the extent of performance also depends on other variables, such as students' anxiety, motivation and metacognitive awareness. Only a limited number of quasi-experimental studies have investigated students' science anxiety, motivation, academic performance and metacognitive development in SRS-supported formative assessments, with most studies being conducted in an university-setting. In addition, studies that use a proper randomized experimental design with control conditions seem to be absent. This dissertation aims to increase our knowledge about providing prompts, such as feedback, on students' science anxiety, motivation, metacognitive awareness, and performance.

The general research question addressed in the studies in this dissertation is:

*‘What types of digital formative assessment improve student learning?’*

In the light of this general research question, this dissertation provides knowledge and insights based on randomized controlled trials (RCTs) in upper secondary physics education, uses empirical analyses, and provides the literature with a new conceptual framework. The contribution of this dissertation is fourfold.

The first contribution is the knowledge and insights it provides to the existing literature and to teachers and instructors in education into the effects of feedback strategies in formative assessments with SRS on students’ science anxiety, motivation, metacognitive awareness, and performance.

The second contribution is the inclusion of potential mediation effects. As far as is known, previous studies have only investigated the direct relationship of an intervention of formative assessments with SRS and an outcome variable of interest. This dissertation also investigates potential mediation effects of (1) anxiety in physics in the relationship between formative assessments and academic performance, and (2) motivation in the relationship between feedback strategies and metacognition, so that teachers and instructors gain insight into the extent to which students’ feelings in a formative assessment affects academic performance and metacognition.

The third contribution of this dissertation is the development of a conceptual framework that identifies relationships between factors that may be responsible for influencing metacognition when students are formatively assessed using SRS. This conceptualization shows relationships between prompts and diagnostic cues that are presented in order to understand how students can improve the accuracy of monitoring judgments and their metacognitive skills when answering questions during formative assessments. The prompts provide teachers with directions as how they can organize a formative assessment, and are useful for students as they help them to identify diagnostic cues that are predictive of subsequent understanding of course content.

The fourth contribution is the large amount of experimental data used. The analyses of these unique data contribute to the existing literature on academic performance, learning gains and metacognition, and these data make it possible to identify what students learn from certain feedback strategies and to what extent these strategies affect students' metacognitive awareness.

## 1.4 Dissertation structure and outline

This dissertation consists of six chapters and is structured in four parts (see Figure 1.1). The two core parts (*Part II* and *Part III*) cover how formative assessments with student response systems (SRS) affect academic performance and metacognition, respectively. The core parts are preceded by *Part I*, which constitutes *Chapter 1* of this dissertation, and describes its aims and contributions, followed by an outline of the dissertation.

*Part II* contains two chapters focusing on academic performance, the first (Chapter 2) dealing with academic performance in a final exam, and the second (Chapter 3) dealing with learning gains while answering questions during formative assessments.

*Chapter 2* describes a randomized experiment in physics courses in upper secondary education that evaluates the effectiveness of formative assessments with SRS on students' performance and students' anxiety. The main research question of this chapter is:

*Do formative assessments improve academic performance and reduce anxiety in physics courses compared to traditional teaching?*

For that purpose, students in a treated condition received formative assessments with SRS, while students in an untreated condition received no formative assessments and did not use SRS. This chapter also examines whether anxiety in physics has a mediating effect in the relationship between formative assessments and students' performance. A series of two-level hierarchical regression models show that formative assessments with SRS reduce anxiety in physics and improve students' performance in a final exam. A mediation analysis proves that anxiety in physics mediates the effects of formative assessments on academic performance, which means that the formative assessments reduce anxiety, which in turn also affects academic performance.

*Chapter 3* presents a randomized field experiment among physics students in upper secondary education, to determine to what extent students learn from teacher feedback, whether or not combined with peer discussions when formatively assessed with SRS. The main research question of this chapter is:

*Do SRS supported assessment activities enhance learning gains?*

The results of the treated students in this chapter are compared with untreated students who receive no feedback, either from teacher or peers. As a measure of what students really learn from answering a question with or without receiving feedback, students in all conditions were asked to individually answer a second question that was similar in difficulty or complexity and assessed the same understanding. Multivariate regressions show that teacher feedback, whether or not combined with peer discussions, positively affects learning gains in comparison with students who receive no feedback. Additional analyses show that student characteristics, time dummies and the level of difficulty of questions do not play a role in explaining the results of learning gain, over and above the differences due to feedback conditions.

*Part III*, consisting of Chapter 4 and Chapter 5, focuses on students' metacognitive awareness when using formative assessments with SRS.

*Chapter 4* is also based on the experiment described in Chapter 3 and evaluates the effectiveness of feedback strategies for the outcome metacognition and the intermediate outcome motivation. The main research question of this chapter is:

*What is the effect of teacher feedback or peer discussions combined with teacher feedback on metacognition and motivation?*

Analysis of variance (ANOVA) shows a positive effect of peer discussions combined with teacher feedback on both metacognitive awareness and motivation compared to the untreated condition. Students with low metacognitive skills benefit more from this treated condition than students with high metacognitive skills. The results show that the effect of peer discussions combined with teacher feedback on metacognition is partly due to its influence on motivation.

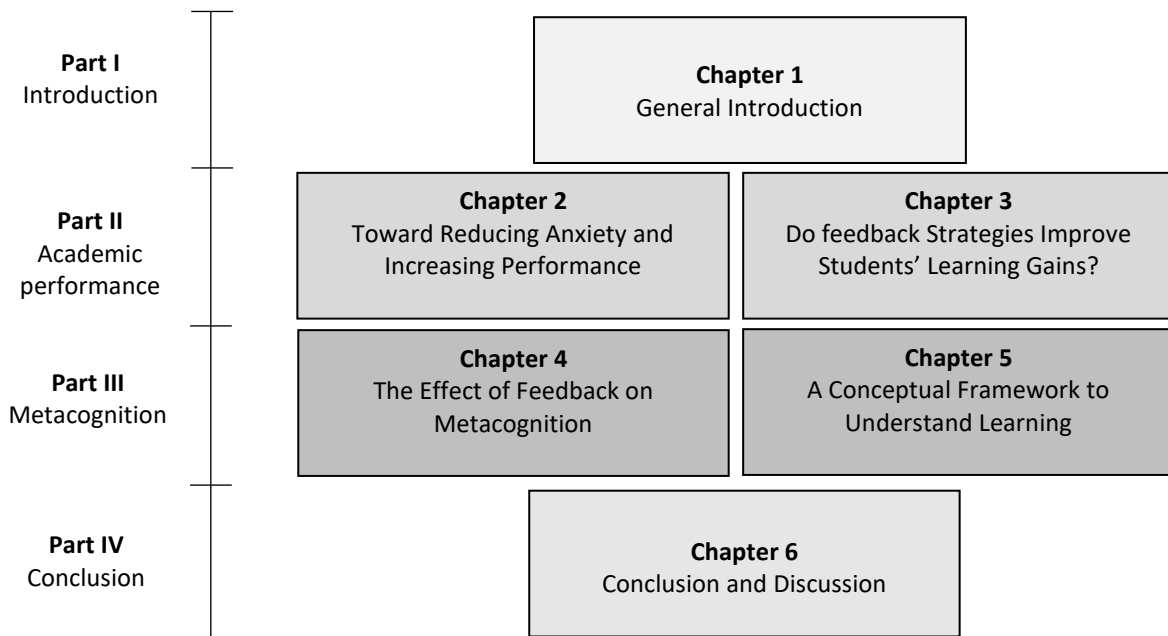
In contrast to the studies described in the previous three chapters, *Chapter 5* is qualitative in nature. The main research question here is:

*How can students improve their accuracy of monitoring judgments and their metacognitive skills when formatively assessed with SRS?*

This chapter presents a new conceptual framework that highlights relationships between factors that may be responsible for affecting metacognitive awareness when students are formatively assessed using SRS. The framework bridges the monitor and control model of Nelson and Narens (1990) and the cue utilization framework of Koriat (1997) and provides insight into how prompts may enhance the utilization of diagnostic cues and thereby increase metacognitive awareness. The framework suggests that more prompts during a formative assessment lead to more diagnostic cues and an enhancement of metacognitive skills.

*Part IV* includes *Chapter 6*. This is the concluding chapter that combines the results of all previous chapters and presents overall conclusions. Finally, this chapter presents and discusses the practical implications, as well as possible future research directions.

Figure 1.1: Contents overview









## Chapter 2

### Toward Reducing Anxiety and Increasing Performance in Physics Education <sup>3</sup>

---

<sup>3</sup> This study is published in Research in Science Education.

Reference: Molin, F., Cabus, S., Haelermans, C., & Groot, W. (2019) Toward reducing anxiety and increasing performance in physics education: Evidence from a randomized experiment. *Research in Science Education*, 51, 233–249.

This chapter evaluates the effectiveness of an intervention of formative assessments with a clicker-based technology on anxiety and academic performance. We use a randomized experiment in physics education in one school in Dutch secondary education. For treated students the formative assessments are operationalized through quizzing at the end of each physics class, where clickers enable students to respond to questions. Untreated students do not receive these assessments and do not use clickers, but apart from that the classes they attend are similar. Findings from multilevel regressions indicate that the formative assessments significantly reduce anxiety in physics, and improve academic performance in physics in comparison with traditional teaching. Furthermore, a mediation effect of anxiety in physics on academic performance is observed. In sum, this implies that an easy to implement technique of formative assessments can make students feel more at ease, which contributes to better educational performance.

## 2.1 Introduction

Anxiety is becoming more and more of a problem in education (Rothman, 2004). Data from the *Program for International Student Assessment (PISA) 2015*, which tests the academic performance of 15-year-olds, showed that more than half of the students feel very anxious before an exam, even if they are well prepared (OECD, 2016). It is not surprising that students experience anxiety before and during exams, because passing or failing a course is often based on a limited number of exams. With so much emphasis on these results, a poor grade in a single exam can have severe consequences, can have a negative effect on a student's final course grade (Burns, 2004) and can even affect admittance to university, course enrolment, career choices, and future employment (Goetz, Bieg, Lüdtke, Pekrun & Hall, 2013; Udo, Ramsey, Reynolds-Alpert & Mallow, 2001). A high level of anxiety can distract students during evaluation and prevent them from recalling relevant information, resulting in a lower than expected performance (Hong, 2010; Maloney, Schaeffer & Beilock, 2013). Anxiety arises when students recognize that their cognitive abilities are overwhelmed by academic demands (González et al., 2017). It is regarded as a serious learning difficulty that hinders students in their learning, students who, by intelligence and hard work, should otherwise perform well (Hong, 2010; Mallow, 2006). Anxiety is even more a problem in science than in other fields, as science-related fields often have specific prerequisites in grades or passing of science courses. Anxiety of science is described as fear of science subjects and science-related situations (Mallow, 1986). It affects students' interest for science lessons and might eventually prevent students from entering certain science-related fields (Hong, 2010; Udo et al., 2004). According to Hong (2010) and Batchelor (2015), only a few studies have examined the effectiveness of interventions aimed at diminishing students' science anxiety. After the review study by Hong (2010), only a limited number of studies has been added to this body of evidence. For example, Brady et al. (2013) and Yu et al. (2014) found that anonymity in classrooms, by student response systems (SRS) where results are not visible to other students, releases students' anxiety and nervousness. Anonymity provides more involvement and active participation of shy or anxious students without peer pressure, while outcomes improve. McDaniel et al. (2011) reported that anxiety reduces, or disappears, when offering students a series of formative assessments or no-stake assessments. By testing course material, students become more familiar with testing and enhance their learning (Roediger & Karpicke, 2006).

These assessments improve students' problem solving skills, without consequences for their grades (Kornell & Son, 2009; McDaniel et al., 2011). Formative assessments are usually accompanied by immediate explanatory feedback. Arkin and Schumann (1984) and Rocklin and Thompson (1985) found that immediate explanatory feedback is associated with less anxiety. Furthermore, by explaining why an answer is correct or incorrect, students will improve their understanding of the course material, which reduces their anxiety and helps them to achieve a higher score in a subsequent attempt (Sullivan, 2017).

## 2.2 Theoretical background

Quizzing is a technique of formative assessment which is easy to implement in classrooms. Formative assessment generated by quizzing can be supported by low-tech or high-tech methods (Fallon & Forrest, 2011). In low-tech classroom settings, students answer multiple-choice questions by, for example, raising their hands, raising coloured flashcards, applause or showing mini whiteboards (Bartsch & Murphy, 2011; Fallon & Forrest, 2011; Wright, Clark & Tiplady, 2018). However, these low-tech tools, although less expensive, have several disadvantages. First of all, it is difficult for a teacher to estimate the number of votes in a limited amount of time, certainly in large classrooms. In addition, due to a lack of anonymity, students can experience pressure of social conformity, and therefore they may not be able to answer honestly (Kay & LeSage, 2009; Stowell & Nelson, 2007).

In high-tech classroom settings, small-handheld electronic student response systems are used. Students can then reply to multiple-choice questions projected on a screen in front of the classroom using SRS, for example, clicker devices (Blasco-Arcas, Buil, Hernández-Ortega & Sesé, 2013; Mayer et al., 2009). Subsequently, answers are collected by IT-software and, depending on the use and purposes of the teacher, the number of correct and false answers can be summarized and displayed on a screen. Collecting information on students' responses in an efficient and fast way can facilitate the teacher to provide immediate feedback, while, at the same time, students remain anonymous.

Previous research has examined the effectiveness of SRS use on improving several aspects of learning, including anxiety. One area of research focuses on general attitudes toward SRS, but also on self-reported anxiety during final exams in classes where these systems are used in

formative assessments. McDaniel et al. (2011) showed that frequent SRS use in science classes reduces anxiety during exams, because SRS stimulate active learning. The immediate feedback of clicker use helps students develop confidence in their science skills (Batchelor, 2015) and improves students' hopes for success (Fallon & Forest, 2011). Agarwal, D'Antonio, Roediger, McDermott & McDaniel (2014) found similar results, because students become familiar with taking formative assessments with SRS and gain insight into which course material matters. Stowell and Nelson (2007) reported that students are more inclined to answer questions in classes where SRS are used. Because students can vote anonymously, teachers create a safe learning environment where students can respond without embarrassment (De Gagne, 2011). SRS can also encourage students to join peer discussions (Yu et al., 2014), so that students may feel less anxious when they first discuss with peers rather than with the teacher before voting anonymously. According to Kay and LeSage (2009), an additional advantage is that students see their own answers positioned in relation to answers of their fellow students, allowing them to monitor their own progress, or get confirmation that they are not alone in their misconception (while still being anonymous).

Another area of research investigates whether academic performance increases in classes where SRS are used, where researchers compare treated conditions (where SRS are used) with untreated conditions (where no SRS are used). In general, these studies found positive effects of improving academic performance by SRS use. Mayer et al. (2009), for example, showed that students score significantly higher in exams when they use SRS to answer two to four questions per lesson. Similar results are also found by McDaniel et al. (2011) and Balta, Perera-Rodríguez and Hervás-Gómez (2018). McDaniel et al. (2011) showed that SRS use in middle school science classes can be extremely effective in increasing academic performance on summative assessments. Balta et al. (2018) demonstrated that the use of *Socrative* in physics classes positively influences students' exam scores.

In sum, the previous literature has shown that formative assessments facilitated with SRS may have the potential to reduce students' anxiety while simultaneously improving academic performance, although one should not overlook the potential drawbacks when implementing SRS. For example, the high initial costs (when clickers are used instead of the free of charge web-based programmes *Socrative* or *Kahoot!*) might be an economic barrier to integrate SRS into class (Blasco-Arcas et al., 2013), while there is always a possibility of technical

malfunctions (Guse & Zobitz, 2011). Furthermore, the extra time required to create multiple-choice questions or to set up these technological devices or programmes may discourage teachers (Lantz & Stawiski, 2014).

The contribution of this study to the literature is threefold. First, to the best of our knowledge, there have been no previous studies that investigated the causal relationship of formative assessments and anxiety in physics, or tested a potential mediation effect of anxiety in physics on the relationship between these assessments and academic performance. Second, most previous studies do not control for selection issues, because there is no randomized trial (e.g. Bachman & Bachman, 2011; Keough, 2012; Shaffer & Collura, 2009), or do not take student characteristics into account (e.g. Bartsch & Murphy, 2011; Fortner-Wood et al., 2013). The few studies that do use an experimental design are not able to distinguish the effect of formative assessments from other effects, such as class attendance in the experiment. For example, in the studies of Mayer et al. (2009) and Morling et al. (2008) students of the treated (SRS) condition are motivated to attend class by earning course credits for answering SRS questions. Students in the untreated (non-SRS) condition do not receive these extra credits for class attendance. In this case, we cannot distinguish whether the estimated improvement in academic performance can be attributed to formative assessments with SRS or to higher class attendance. In the study of this chapter, we carry out a randomized experiment that solely focuses on the effectiveness of formative assessments with SRS and do not use reward systems for students' attendances or students' responses, or other potential confounding treatments. Third, we study the effect of formative assessments over a traditional teaching approach in secondary school, whereas the systematic literature review of Kay and LeSage (2009) showed that most of this research was done at university-level. Only a few studies have analyzed the effectiveness in secondary education (Vital, 2011). Nevertheless, this knowledge is deemed necessary, since SRS usage is rapidly increasing in secondary education.

The effectiveness of formative assessments with SRS is evaluated for two particular outcomes, namely anxiety and academic performance. Our research questions are:

1. Do formative assessments reduce anxiety in physics compared to traditional teaching?

2. Do formative assessments improve academic performance in physics compared to traditional teaching?
3. Does anxiety work as a mediating factor for the effect of formative assessments on academic performance?

For this purpose, we have used a randomized experiment to examine the effects of formative assessments with SRS on physics performance and anxiety in physics in secondary education, taking the potential drawbacks into account. For its implementation in daily teaching practice, the formative assessments in the treated condition are facilitated with clickers, while the untreated condition does not use in-class questioning or clickers, but instead, follows traditional physics teaching.

In the sequel of this chapter we will only use the term formative assessments when referring to formative assessments with SRS (in this case clicker) use.

## 2.3 Methods

### 2.3.1 The intervention

The intervention consists of formative assessments, in which we used a method similar to Mayer et al. (2009). There, students in the treated condition used clickers to answer two to four multiple-choice questions per lesson, while identical students in the untreated condition did not use in-class questioning or clickers. In our study, we similarly apply clickers for formative assessment in class, although it should be noted that we do not use reward systems for students' responses, in contrast with the study of Mayer et al. (2009).

In our study, a total of 73 treated students used a clicker-supported questioning method as a form of formative assessment, and 66 untreated students did not use in-class questioning or clickers, but, instead, followed traditional physics teaching. The treated condition and the untreated condition of each education level were both taught by the same teacher (Figure 2.1), to minimize potential teacher effects that might otherwise influence the results. The three physics teachers collaborated voluntarily in this study. In line with other studies (e.g. Mayer et al., 2009; Pan et al., 2019), the teachers taught both conditions identically: students of each educational level received the same lecture contents, notes, assignments, and exam questions. The main (and only) difference between the two sections was the way



the teacher interacted with the students at the end of each lesson. Three times per week, at the end of each lesson, students in the treated condition were formatively tested for about 10 to 15 minutes for a period of 17 weeks. In particular, the clicker was used by treated students to answer four multiple-choice questions each lesson. Each question covered a part of the course content, which provided the teacher with valuable information about students' understanding, and the students an insight into the level of their comprehension of the course material (Beatty, Gerace, Leonard & Dufresne, 2006). The multiple-choice questions had four possible answers, while the fifth option was "I don't know". This latter option should minimize guessing of students and inform the teacher when students really did not know the answer to the question (Caldwell, 2007).

Figure 2.1: Course and design of the intervention

Treated condition	73 students			66 students			Untreated condition
Physics teaching <u>AND</u> 10-15 minutes formative assessment with a clicker- supported questioning method. (3 times a week)	<b>Class 1</b> 10 <sup>th</sup> grade		<b>Teacher 1</b>	<b>Class 2</b> general secondary education		Physics teaching with in class homework time and feedback opportunity (but <u>WITHOUT</u> in-class formative assessment). (3 times a week)	
	<b>Class 3</b> 10 <sup>th</sup> grade		<b>Teacher 2</b>	<b>Class 4</b> pre-university education			
	<b>Class 5</b> 11 <sup>th</sup> grade		<b>Teacher 3</b>	<b>Class 6</b> pre-university education			

All multiple-choice questions were inserted into *PowerPoint* slides and projected on a screen in front of the class. In line with the difficulty of the questions, students were given limited time for considering the question individually or discussing with their peers, after which they answered the question individually by choosing the corresponding button on their clicker. The responses of all clickers were registered and recorded by the 'TurningPoint Technologies' software. After each round of answering a multiple-choice question, the software presented the distribution of answers as a bar graph to all students on the screen in front of the class. Next, they heard the correct answer and received the teacher's feedback on the most common mistakes made by students. Depending on the variety of answers, the teacher could decide whether to spend limited or more elaborated time on feedback. On average, students

spent around 1 minute per question to provide an answer and around 2 minutes per question receiving feedback from the teacher on their answers. The purpose of these sessions was not only to provide feedback to students about common misunderstandings or misconceptions, but also to evaluate students' understanding of the course material and to visualize academic progression (Premuroso, Tong & Beed, 2011). The answers of the treated students on the multiple-choice questions were not graded. In fact, the responses to the questions were completely anonymous, so that there were no drawbacks for students to answer the questions.

The untreated students did not receive in-class questioning using clickers at the end of each lesson. Instead, they followed a traditional teaching approach by completing their homework independently or with peers, where they had the opportunity to ask questions of the teacher, and received additional feedback via that way. Figure 2.1 gives a visual representation of the intervention.

Note that the intervention of formative assessments consisted of an inseparable combination of answering multiple-choice questions and feedback on the responses to these questions. Researchers have compared various ways to implement formative assessments, resulting in fairly detailed instructions about effective instructional design on the matter, e.g. regarding feedback (Larsen & Butler, 2013). In this regard, it has been shown that feedback is most effective if given immediately, and if it is substantively elaborate (Roediger & Butler, 2011). The current consensus in the literature is that the learning process is more effective when students receive feedback on their progress during the instruction period. For this reason, we chose to provide feedback to students immediately after displaying the histogram. Moreover, the literature stresses that the effect of formative testing may not be limited to its information value ('feedback effect') and related learning incentive. Testing also contributes to learning by itself (Roediger & Karpicke, 2006).

One potential threat for the design of the intervention is that treated and untreated students can share the didactic material. This would violate the independence assumption. In order to prevent students from sharing the exact multiple-choice questions with untreated students, students of the treated condition made their notes with pen and paper and left these in the classroom at the end of each lesson. In addition, they were not allowed to use their mobile

phones during lessons, in order to avoid students taking pictures. Here, the supervision of the teacher was crucial to make sure that no learning material of the multiple-choice questions was taken home. If the independence assumption is violated, and untreated students would have had the same information of the multiple-choice questions, it is likely that this would undermine the true effect.

### **2.3.2 Assignment to the treatment**

The experiment was conducted in one school in the southern part of the Netherlands. The school is a typical, average mid-sized school outside the highly urbanized, central region of the Netherlands, offering secondary education to 1,500 students at three ability levels: theoretical pre-vocational education (four years), general secondary education (five years), and pre-university education (six years). The participating school had been invited by contacting the principal by email. In this email, the purpose and set-up of the study were explained. In a follow-up personal meeting, the researchers explained the importance of randomization of the participants and suggested randomizing before the timetable was made, to prevent timetabling issues interfering with the possibility of randomization. The principal was convinced by the importance of the study and reassured that the study would not interfere with other issues at school. Moreover, the study was unlikely to harm the students or their performance and so the principal agreed. After agreement, the involved teachers were informed about the aims of the study.

One hundred and thirty-nine physics students participated in the study over a period of 17 weeks at the start of the school year 2016-2017. The students belonged to six different classes; two of 10<sup>th</sup> grade general secondary education, two of 10<sup>th</sup> grade pre-university education and two of 11<sup>th</sup> grade pre-university education. All the participants were first randomly assigned by the scheduling software *Zermelo* to one of two classrooms of each education level, after which the school timetable was made. Next, the classes of each education level were allocated randomly to a treated or untreated condition by the researchers. This assignment procedure successfully constructed a comparable treated- and untreated condition and accounted for potential selection-into-treatment effects, as will be discussed in the next section.

### 2.3.3 Empirical strategy

We use a series of two-level hierarchical regression models to test whether formative assessments affect academic performance and anxiety. This multilevel analysis can be formulated as follows:

$$Y_{ij} = \alpha + \beta_0 \theta_{ij} + \sum_{k=1}^k \beta_k x_{k,ij} + (\mu_j + \varepsilon_i), \quad (2.1)$$

where  $Y_{ij}$  denotes the outcome variable anxiety in physics or academic performance in a final exam of student  $i \in \{1, 2, \dots, 139\}$  attending physics class  $j \in \{1, 2, \dots, 6\}$ ;  $\theta_{ij}$  the intervention dummy (0,1) with 0 for the untreated condition and 1 for the treated condition;  $X_{ij}$  a vector of observed student characteristics prior to the experiment; and  $\varepsilon_i$  the standard error (note we include control variables to increase the precision in estimation of the effects of our intervention on differences in the outcome measures between students). We also included  $\mu_j$ , a parameter which denotes unobserved information at the class level. Previous literature has pointed out that teachers' teaching style and interactions between students in class are important factors that may influence the results of the students, regardless of the intervention (Chetty, Friedman & Rockoff, 2011; Koth, Bradshaw & Leaf, 2008). Because we are interested in the causal impact of the treatment on students' outcomes  $Y_{ij}$ , it is not desirable that teachers are influencing student outcomes. Therefore, one class of each teacher is randomly assigned to the untreated condition and the other to the treated condition. Furthermore, the intervention takes place in several classes, so there is an individual learning process and a group learning process. The group learning process is an important outcome of the intervention. Because treated students were allowed to have peer discussions on the multiple-choice questions, it is possible that interactions between students in the classes arose. These interactions may also play a significant role in students' anxiety in physics (Guarascio, Nemecek & Zimmerman, 2017). Therefore, we introduce a multilevel random effects model in order to control for unobserved class (and thereby teacher) level variance in the regression.

Previous literature indicates that formative assessments may not only improve academic performance, but may also reduce anxiety. This immediately raises the question of whether the relationship between these assessments and academic performance are mediated by

anxiety in physics. A mediation effect will be assessed using Baron and Kenny's (1986) test for mediation and is present as (1) the intervention significantly impacts academic performance; (2) the intervention has a significant effect on the presumed mediator anxiety in physics; (3) the presumed mediator anxiety in physics is significantly associated with academic performance; and (4) the intervention is no longer significant (complete mediation), or is reduced (partial mediation), when the post-test score of anxiety in physics is included in a model that tests the causal effect of the intervention on academic performance.

To reduce the risk of type I statistical error, for all estimates in our multilevel random effects models we use Bonferroni's adjustment alpha level of 0.025 (Grove & Andreasen, 1982).

### **2.3.4 Outcome measures**

#### *2.3.4.1 Pre-treatment information*

While a number of SRS-related studies only compare post-outcomes of participants in treated conditions and untreated conditions (Hunsu, Adesope & Bayly, 2016), this study also uses pre-treatment characteristics on demographic, academic, and non-cognitive performance skills of the students.

First, we collected demographic data and pre-treatment physics grades of all (treated and untreated) students who participated in the intervention from administrative data of the school (Figure 2.2). The physics grades are computed as an average of all exam grades in the school year before the intervention. In the Dutch education system, the traditional grading scale is a 90-point scale from 1.0 through to 10.0, with 5.5 being the minimal pass grade. This scale is subdivided with intervals of one decimal place. All physics grades are converted to z-scores in order to calculate effect sizes. By standardizing, we ensure that a reader who is not familiar with the Dutch grade system is also able to interpret the effects.

Furthermore, we collected data on five non-cognitive components, namely: motivation (extent to which students like studying), concentration (extent to which students can concentrate during homework), study approach (extent to which students study efficiently), task approach (extent to which students tackle study components systematically), and memory (extent to which students learn until they know 'everything') (Table 2.1). To measure these components, we used a validated, self-reported questionnaire called '*Study Conditions*

*Questionnaire*' (*Vragenlijst Studievoorwaarden*) (Crins, 2002) consisting of 38 three-point Likert scale items ranging from never to always. In further analyses we use a standardized (z-)score of each of the values on the components in order to facilitate the interpretation of results. Higher z-scores indicated more motivation, a better concentration, a better study approach, a better task approach and more memorization. The internal consistency reliabilities, as measured with Cronbach's alpha, for all of the five components were acceptable (Tavakol & Dennick, 2011): motivation:  $\alpha = 0.65$ , concentration:  $\alpha = 0.80$ , study approach:  $\alpha = 0.68$ , task approach:  $\alpha = 0.73$ , and memory:  $\alpha = 0.79$ .

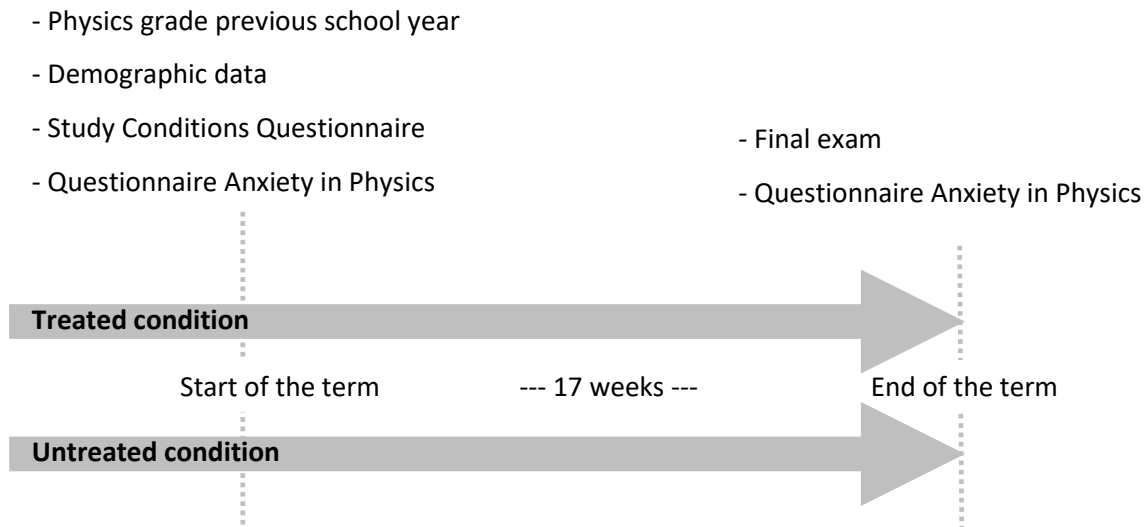
For the pre-treatment outcome variable 'anxiety in physics', we collected data from the students using the *Mathematics Anxiety Scale* of Betz (1978) adjusted for physics. The scale was intended to assess feelings of anxiety and nervousness related to doing physics and consisted of 10 four-point Likert scale items ranging from 'strongly disagree' to 'strongly agree'. Positively worded items were reversed, so that higher scores indicated more anxiety in physics. The scores on this scale were converted to z-scores (Table 2.1). The internal consistency reliability of this anxiety scale was  $\alpha = 0.68$ .

#### 2.3.4.2 Post-Treatment information

In both treated and untreated conditions the academic performance was measured after 17 weeks in a final exam (score out of 50 points). This exam incorporated none of the multiple-choice questions from the clicker sessions, but consisted solely of open questions of standardized Dutch national physics exams constructed by the *Central Institute for Test Development* (the Dutch abbreviation for this organization is CITO). Each teacher marked the exams of their colleague according to a uniform correction model, not knowing which students belonged to the treated condition and the untreated condition. Hereby, we pursued an increased reliability and internal validity of the test scores on the exam. The scores on these exams were converted to z-scores.

For the measure of anxiety in physics, the students completed the same questionnaire '*Mathematics Anxiety Scale*' adjusted for physics at the end of the intervention (Figure 2.2). The internal consistency reliability of this anxiety scale was  $\alpha = 0.79$ .

Figure 2.2: Research design



### 2.3.5 Descriptive statistics

In total, 139 students from 6 different classes participated in this study. At the start of the intervention, the students were on average 16.2 years old ( $SD = 0.69$ ). Fifty-two percent of the participants were male, while students had an average physics grade of 6.66 points ( $SD = 1.02$ ) in exams in the previous school year.

Table 2.1 presents a comparison of the observable characteristics of the treated condition and the untreated condition, as well as the statistics of the (significance of the) mean differences. The quality of the randomization was examined using independent two sample  $t$ -tests. The independent two sample  $t$ -test shows that students in the treated condition and untreated condition scored, on average, the same on physics exams. However, the score for study approach significantly differed between the treated condition and untreated condition; students in the treated condition studied more regularly and more efficiently. Except for the variable study approach, students of the treated condition were, on average, similar with students of the untreated condition. Also, a joint  $F$ -test on all the characteristics does not show a significant difference;  $F(9, 129) = 1.09, p = 0.37$ .

Table 2.1: Mean differences on standardized pre-treatment characteristics

	<i>Treated condition</i> ( <i>N</i> = 73)		<i>Untreated condition</i> ( <i>N</i> = 66)		<i>Diff</i>	<i>p</i>
	<i>Mean</i>	<i>Std. Dev.</i>	<i>Mean</i>	<i>Std. Dev.</i>		
Gender <sup>A,B</sup>	0.42	0.50	0.55	0.50	- 0.13	0.16
Age <sup>A</sup>	- 0.023	0.12	0.025	0.12	- 0.048	0.77
Physics grade <sup>A</sup>	0.055	0.12	- 0.061	0.12	0.12	0.50
Anxiety pre-score <sup>A</sup>	0.0063	0.12	- 0.0069	0.12	0.013	0.18
Motivation <sup>A</sup>	0.12	0.12	- 0.13	0.12	0.25	0.14
Concentration <sup>A</sup>	0.11	0.11	- 0.12	0.13	0.23	0.18
Study approach <sup>A</sup>	0.20	0.11	- 0.22	0.13	0.42	0.014**
Task approach <sup>A</sup>	0.093	0.13	- 0.10	0.11	0.19	0.25
Memory <sup>A</sup>	0.048	0.12	- 0.053	0.12	0.10	0.55

\*\*\* = significant at the 1 % level; \*\* = significant at the 5 % level; \* = significant at the 10 % level.

<sup>A</sup> standardized scores

<sup>B</sup> boy = 0, girl = 1

- Measured at T0 for treated and untreated condition.

In the next section, we will control for all of these student background variables in multilevel regressions, when we analyze if formative assessments with SRS affect anxiety in physics and improve academic performance.

## 2.4 Results

### 2.4.1 Anxiety in physics

The first two models of Table 2.2 present the results of the multilevel analyses with the outcome measure ‘anxiety in physics’. Model 1 only includes the intervention dummy. Next, we gradually add observed pre-treatment characteristics. The results of Model 1 indicate a positive effect of the intervention on anxiety in physics with  $\hat{\theta}$  equal to - 0.37 points of one standard deviation, significant at 2.5 % level (i.e. with Bonferroni correction applied). This corresponds to a small to medium effect size (Cohen, 2013) and indicates that the intervention significantly reduced anxiety in physics. In the second model, the control variables gender, physics grade of previous year, anxiety pre-score and a set of five non-cognitive variables are added. Adding these pre-treatment control variables in the analysis increases the precision of our estimates, as they can predict the differences in the outcome (Bloom, Richburg-Hayes & Black, 2007; Raudenbush, 1997) and thereby make the model more effective. Even though most control variables were not significantly different between treated



and untreated students before the intervention, they still influence the outcome and thereby add precision to the second model. Model 2 indicates that the effect of the intervention on academic performance ( $\hat{\theta} = -0.28$ ) remains significant at 2.5 % level. Furthermore, gender ( $\hat{\beta}_1 = 0.36, p < 0.01$ ), physics grade of last year ( $\hat{\beta}_2 = -0.27, p < 0.01$ ), anxiety pre-score ( $\hat{\beta}_3 = 0.37, p < 0.01$ ) and motivation ( $\hat{\beta}_5 = -0.30, p < 0.01$ ) are significant predictors of the post-anxiety score. This means that students who already experienced more anxiety before the intervention, also experience more anxiety in physics after the intervention. On the other hand, the post-score anxiety is lower if students are more inclined to study or willing to commit themselves to their study (variable motivation). The variable study approach, in which the treated condition and untreated condition differed significantly before the intervention, does not significantly predict the post-score anxiety, nor does it influence the significance and magnitude of the intervention.

#### 2.4.2 Academic performance

Models 3 and 4 of Table 2.2 present the results for the outcome variable academic performance. In Model 3, the estimate of  $\hat{\theta}$  is equal to 0.46 points of standard deviations significant at the 1 % level and corresponds to a medium effect size (Cohen, 1988). When additional pre-treatment control variables are included in Model 4, the effect of participation in the treated condition remains robust with  $\hat{\theta}$  equal to 0.34 points of standard deviations, significant at 2.5 % level. Furthermore, gender ( $\hat{\beta}_1 = -0.30, p < 0.05$ ), physics grade of last year ( $\hat{\beta}_2 = 0.57, p < 0.01$ ) and motivation ( $\hat{\beta}_5 = 0.24, p < 0.01$ ) are significant predictors of academic performance. After controlling for a couple of other factors, we see that boys still perform better than girls in physics exams. Therefore, we also estimated our model with interaction effects between intervention and gender, to see if the intervention might have a differentiating effect by gender, but we did not find a significant effect.

Table 2.2: Multilevel regression analyses predicting post-score anxiety and academic performance

	Model 1 <i>Anxiety post-score</i>	Model 2 <i>Anxiety post-score</i>	Model 3 <i>Academic performance</i>	Model 4 <i>Academic performance</i>	Model 5 <sup>A</sup> <i>Academic performance</i>	Model 6 <sup>A</sup> <i>Academic performance</i>
( $\theta$ ) Intervention	- 0.37** (0.17)	- 0.28** (0.13)	0.46*** (0.16)	0.34** (0.17)	--	0.24 (0.16)
( $\beta_1$ ) Gender (boy = 0 & girl = 1)	--	0.36*** (0.13)	--	- 0.30* (0.13)	- 0.21 (0.13)	- 0.19 (0.13)
( $\beta_2$ ) Physics grade	--	- 0.27*** (0.066)	--	0.57*** (0.066)	0.48*** (0.067)	0.47*** (0.066)
( $\beta_3$ ) Anxiety pre-score	--	0.37*** (0.069)	--	- 0.0032 (0.067)	0.12 (0.070)	0.11 (0.070)
( $\beta_4$ ) Study approach	--	0.12 (0.097)	--	- 0.057 (0.095)	- 0.016 (0.089)	- 0.018 (0.090)
( $\beta_5$ ) Motivation	--	- 0.30*** (0.094)	--	0.24*** (0.092)	0.15 (0.089)	0.15 (0.089)
( $\beta_6$ ) Concentration	--	- 0.089 (0.086)	--	- 0.012 (0.083)	- 0.034 (0.079)	- 0.041 (0.079)
( $\beta_7$ ) Task approach	--	0.040 (0.086)	--	- 0.022 (0.085)	- 0.016 (0.080)	- 0.013 (0.080)
( $\beta_8$ ) Memory	--	0.17 (0.069)	--	- 0.060 (0.086)	- 0.0067 (0.082)	- 0.0065 (0.082)
( $\rho$ ) Anxiety post-score	--	--	--	--	- 0.34*** (0.078)	- 0.33*** (0.078)
Constant	0.20 (0.12)	2.15*** (0.44)	- 0.24** (0.12)	- 4.11*** (0.44)	- 3.31*** (0.45)	- 3.38*** (0.45)

Table 2.2 (continued)

	Model 1 <i>Anxiety post-score</i>	Model 2 <i>Anxiety post-score</i>	Model 3 <i>Academic performance</i>	Model 4 <i>Academic performance</i>	Model 5 <sup>A</sup> <i>Academic performance</i>	Model 6 <sup>A</sup> <i>Academic performance</i>
<b>Random parameters</b>						
Student level variance	0.96	0.54	0.94	0.52	0.46	0.46
Class level variance	0.00	0.00	0.00	0.018	0.034	0.017
Deviance (-2 log likelihood)	388.55	309.58	385.84	307.34	292.68	290.84
Observations	139	139	139	139	139	139
$\rho$	0.00	0.00	0.00	0.034	0.068	0.035

\*\*\* = significant at the 1 % level; \*\* = significant at the 2.5 % level; \* = significant at the 5 % level.

<sup>A</sup> Mediation analysis predicting academic performance (Model 5 and Model 6).

- Control variables gender, physics grade previous school year, anxiety pre-score, study approach, motivation, concentration, task approach and memory has been measured at T0. Anxiety post-score has been measured at T1.

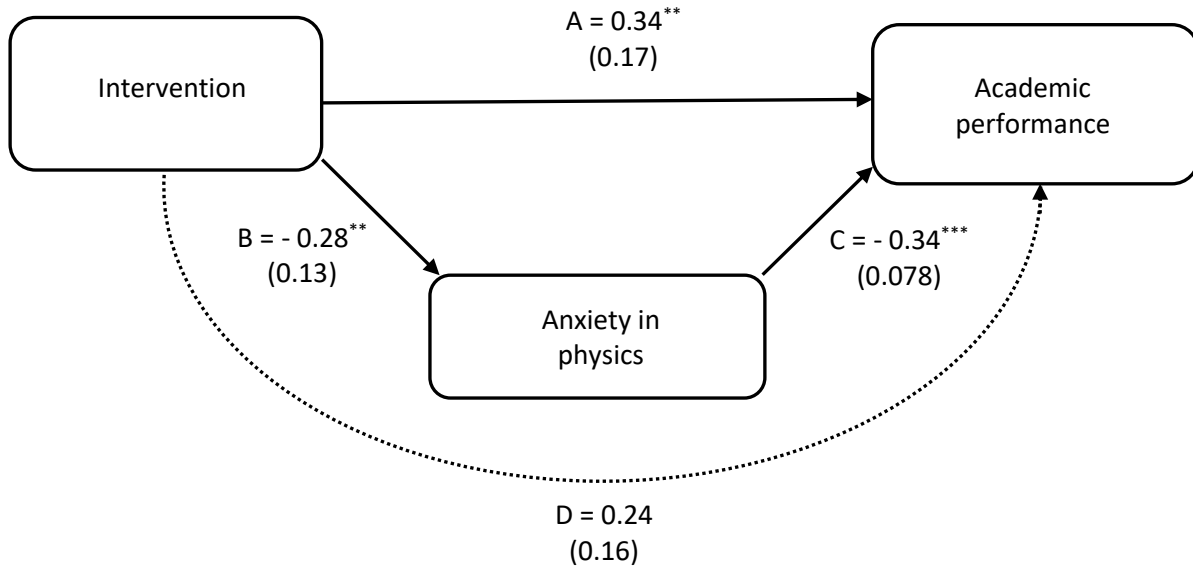
### 2.4.3 Mediation analysis

Anxiety in physics might mediate the estimated effects of the intervention on academic performance. Baron and Kenny (1986) proposed four steps for testing this kind of mediation effect. First, consider again the multilevel analysis showing that the intervention had a significant positive effect on academic performance (Model 4 in Table 2.2;  $\hat{\theta} = 0.34, p < 0.025$ ). This effect is presented in Figure 2.3, path A. Second, consider the multilevel analysis showing that the intervention had a significant negative effect on anxiety in physics (Model 2 in Table 2.2;  $\hat{\theta} = -0.28, p < 0.025$ ). This effect is presented in path B in Figure 2.3. Then, third, the post-score on anxiety in physics is included into a regression with outcome academic performance. If not controlled for the intervention dummy, the results indicate that the presumed mediator anxiety in physics significantly correlates to academic performance (Model 5 in Table 2.2;  $\hat{\rho} = -0.34, p < 0.01$ ). This effect is presented in Figure 2.3, path C. It is now intuitive that anxiety in physics might mediate the estimated effect of the intervention on academic performance, and that this mediation effect can be revealed by including the post-scores on anxiety in physics into the regression (Table 2.2; Model 6). Doing so, however, shows that the estimate of  $\hat{\theta}$  no longer significantly predicts academic performance;  $\hat{\theta} = 0.24, ns$  (path D in Figure 2.3). Therefore, it is concluded that anxiety in physics completely mediates the effects of the intervention on academic performance. Apart from the insignificant coefficient, the decrease in effect size (from 0.34 to 0.24) is an indication that the effect of the intervention on performance is mediated by anxiety in physics.

To conclude, we also calculated the intraclass correlation coefficients (ICC) to estimate the percentage of variance of the outcomes anxiety in physics or academic performance explained by unobserved class effects. In most models, the ICC's are quite low, varying from  $\rho = 0.00$  to  $\rho = 0.07$ . This means that almost all the variance is explained by student differences and not by unobserved class effects (Peugh, 2010). However, although in most models, we find that the percentage share of variance of the outcome variables explained by unobserved class effects is less than 0.05 (which is the rule of thumb from Hox (1998) for deciding against the use of the multilevel model), we still opt for the multilevel random effects model. The most important reason is that in model 5, and to a lesser extent in models 4 and 6, we do observe class effects. Furthermore, the multilevel model also allows us to account for class differences (McNeish, 2014), whereas using fixed effects for class (which also accounts for class

differences) adds coefficients to the models, which lowers the degrees of freedom and is not a good idea given our number of observations.

Figure 2.3: Mediation analysis



\*\*\* = significant at the 1 % level; \*\* = significant at the 2.5 % level; \* = significant at the 5 % level.

## 2.5 Conclusion and discussion

The aim of this chapter was to evaluate the effects of using repeated formative assessments compared to traditional teaching. A randomized experiment was carried out over a period of 17 weeks among 139 secondary students and analyzed at student level, while controlling for class in a multilevel setting. The study answers our three research questions. First, the results show that formative assessments improve academic performance in physics compared to traditional teaching; treated students have significantly higher grades on the post-test than untreated students. It corresponds to an effect of 0.34 points of standard deviations, significant at 2.5 % level, a medium effect size. This finding is in line with the studies of e.g. Bachman and Bachman (2011), Lin, Liu and Chu (2011), and Mayer et al. (2009), who found that students in a clicker group outperform students in a non-clicker group. It should be noted that the untreated condition in our study had more time and more opportunities to ask questions of the teacher. The difference in additional time and more opportunities to ask questions to the teacher apparently did not help the untreated condition. Second, compared to traditional teaching, repeated formative assessments significantly reduce anxiety in

physics. This effect is equal to 0.28 points of a standard deviation, significant at 2.5 % level, a small to medium effect size. These findings differ from the findings in the scarce studies in the literature on this topic; Sun (2014) and Bachelor (2015) both showed insignificant differences in anxiety between treated classes and untreated classes, although both studies suffer from a low power. Noteworthy is that the study of Bachelor (2015) measured a statistical increase in anxiety during the semester in both classes. However, given that these studies focused on university settings, it is hard to compare these effects with secondary education. And third, a mediation analysis shows that anxiety in physics completely mediates the effects of formative assessments on academic performance. This means that this form of assessment significantly reduces anxiety, which in turn also significantly affects academic performance. Although we may have expected the class level to introduce bias in the results (since peer discussions between students may play a significant role in improving academic performance and reducing students' anxiety (Wiggs, 2011)), the analysis does not show significant bias, neither on anxiety in physics nor on academic performance. We did not collect information from students that could explain this, but the teachers indicated that most of the treated students answered multiple choice questions individually and in silence, without consulting their peers. As a result of these findings, we are inclined to assume that treated students experience less anxiety and become more familiar with taking assessments, because they receive more instances of performance feedback in a single lesson, than untreated students. The repeated assessments divide the course material into small units and the real-time feedback from the teacher helps students to monitor their own understanding of the material. Simultaneously, the anonymity enabled by clickers and the general group feedback is less risky to expose students' weakness to other students and teacher, which could give rise to anxiety. On the other hand, anxiety in physics of untreated students is not addressed specifically, because they do not receive in-class questioning and receive only selective feedback when asked for it. These assumptions are in line with the literature that a series of anonymous formative assessments with multiple-choice questions and explanatory group-feedback of teachers enables students to increase their self-confidence and improves their metacognitive skills that reduces anxiety (Brown, Roediger & McDaniel, 2014; Kornell & Son, 2009; Sullivan, 2017).

Despite our diligent preparations, at least two caveats should be considered when interpreting the results. First, the students, who participated in this study, had not used clickers before. It is therefore possible that students were excited about using new course material, were studying harder, or even had more commitment to the course material, than before the introduction of clickers. As such, due to the technology, the estimated effect of the intervention may be upwardly biased. We did not collect information that could unravel novelty effects; however, teachers have said that students became more accustomed to using clickers over time, accepting clickers as a standard study tool. Since this study lasted 17 weeks, we expect that novelty effects faded out over time and did not, or at least not substantially, bias our results. Second, the scope of this study was limited to physics teaching at only one secondary school in the Netherlands. Further research should indicate whether the findings in this study also apply to other science subjects taught in secondary education.

In conclusion, although the intervention was carried out among only 139 students, it is important to indicate that substantial effects can be achieved in secondary school physics courses if students are formatively assessed for only 10 to 15 minutes each lesson. These positive outcomes may stimulate physics teachers to implement formative assessments with clickers in their lessons.







## Chapter 3

### Do Feedback Strategies Improve Students' Learning Gains? <sup>4</sup>

---

<sup>4</sup> This study is published in *Computers & Education*.

Reference: Molin, F., Haelermans, C., Cabus, S., & Groot, W. (2021). Do feedback strategies improve students' learning gain? - Results of a randomized experiment using polling technology in physics classrooms. *Computers & Education*, 104339.

In this chapter we study if feedback strategies during formative assessment with polling technology have an impact on learning gains in the short term. We conduct a randomized experiment in physics class in upper secondary education with the web-based student response system (SRS) *Socrative* comparing three conditions. In the cooperative condition students receive a combination of peer discussion and teacher feedback, while in the individual condition they receive only teacher feedback. In the control condition students do not receive any feedback, from either teacher or peers. To measure what individuals learn from teacher feedback, whether or not combined with peer discussions, students in all three conditions individually answer paired isomorphic multiple-choice questions. Per question, we study the probability to answer the second isomorphic question correct and compare this between conditions. The analyses show that teacher feedback, regardless of the peer discussions, positively affects learning gains in question pairs of the treatment conditions in comparison with the control condition. However, the cooperative condition shows that the largest learning gains occur when peer discussion is followed by teacher feedback. The findings provide insights into which feedback strategies affect learning gains in question pairs and are important for teachers when implementing feedback strategies during formative assessments.

### 3.1 Introduction

Feedback is the core of formative assessment and one of the most important aspects of students' learning (Black & Wiliam, 1998a; Hattie & Timperley, 2007). It is seen as an ongoing process in which students in dialogue with their teacher and peers are encouraged to monitor, evaluate and regulate their own learning, to develop their abilities as self-regulated, metacognitively aware learners (Boud & Molloy, 2013; Evans, 2013). Feedback provided by teachers or peers that serves to modify the teaching and learning activity is regarded as the bedrock of formative assessment (Black & Wiliam, 1998a). Students receive feedback from teachers or peers on their understanding to make informed decisions in their next step of learning, while teachers receive feedback from students in order to reflect on their teaching and peers receive feedback from other students to detect and correct errors in their reasoning. Formative assessments raise students' understanding, enhance learning gains and improve metacognitive awareness and performance (Hattie & Timperley, 2007; Nicol & Macfarlane-Dick, 2006; Sadler, 1998). It is sometimes not easy for teachers to organize formative assessments, because they do not know which of the students understands the course content (Dudaitè & Prakapas, 2017). Real-time data collected by student response systems (SRS) or polling technologies can assist teachers in tailoring feedback (Chien, Lee, Li & Chang, 2015). Formative assessments can be easily organized by SRS<sup>5</sup> that provide students and teachers with instant feedback (Hwang & Tsai, 2011; Krumsvik & Ludvigsen, 2012; Ludvigsen, Krumsvik & Furnes, 2015; McDonough & Foote, 2015; Shapiro et al., 2017; Sung, Chang & Liu, 2016). In these systems, students answer multiple-choice questions posed by their teacher by pressing a button on their clicker device, smartphone or similar web-based response system, after which they can receive feedback from the system, their peers or teacher. The effectiveness of a formative assessment on learning gains will depend, among other things, on the chosen feedback strategy. The effectiveness -measured by learning gains- of specific feedback strategies has been studied extensively. Several studies have shown that

---

<sup>5</sup> SRS such as *Socrative*, *Kahoot!* and *TurningPoint* are widely used in schools, but how these systems can be used strategically, how they affect teaching, and what problems need to be addressed in this process remain mostly unanswered (Becker et al., 2016). In fact, schools usually do not use technology effectively for learning or teaching (Becker et al., 2016; Johnson et al., 2016). The use of technology in the classroom is often an unfortunate side effect of meeting budgetary and other time-sensitive deadlines without sufficient attention for any kind of research that demonstrates effectiveness (Brady, Rosenthal, Forest & Hocevar, 2020).

the number of multiple-choice questions answered correctly increases as students re-vote the same question after a peer discussion (Crouch & Mazur, 2001; Egelandstad & Krumsvik, 2017a; Mazur, 1997; Porter, Bailey Lee, Simon & Zingaro, 2011; Smith et al., 2009; Smith, Wood, Krauter & Knight, 2011; Zingaro & Porter, 2014). Egelandstad and Krumsvik (2017), Porter et al. (2011), Smith et al. (2009) and Zingaro and Porter (2014) all reported that students learn during peer discussions and that they do not merely copy answers from their peers. Smith et al. (2009) and Egelandstad and Krumsvik (2017) showed that a combination of peer discussion followed by teacher feedback leads to substantially greater learning gains when compared to peer discussions or teacher feedback alone. Smith et al. (2009) and Porter et al. (2011) also both found that students show the largest learning gains when they first engage in a peer discussion and then listen to an explanation by the teacher.

However, most studies do not control for selection issues, as they are not based upon randomized experiments. One of the contributions of the study at hand is that we set up a randomized experiment, allowing us to attribute effects. Furthermore, our study is not only limited to a condition in which students only receive teacher feedback combined with peer conditions (as in the studies of e.g. Egelandstad and Krumsvik (2017), Porter et al. (2011), Smith et al. (2009) and Zingaro and Porter (2014)), but also includes a condition in which students only receive teacher feedback. By comparing these two treatment conditions with a control condition, we can draw strong conclusions about the contribution of a specific feedback strategy to learning gains. Finally, while most studies are based on small samples in (mostly) university settings with a limited number of multiple-choice questions, we use results from a large randomized experiment in upper secondary education.

The main contribution of this study is that we address whether SRS supported assessment activities impact learning gains. We provide insight into feedback strategies and learning gains and compare learning gains in question pairs of treatment conditions with the control condition. The research questions that guide the study are:

1. Do technology/polling system supported assessment activities enhance learning gains?
2. Are learning gains modified by peer discussions?

To this end, we have conducted a randomized experiment in which we tested three conditions in physics education in upper secondary education in the Netherlands. Each week, students answered conceptual multiple-choice questions individually with the web-based response system *Socrative*. Students in the individual condition received immediate teacher feedback on how well a concept of the question is understood, while students in the cooperative condition discussed this question with their peers and answered it for a second time before receiving feedback from the teacher. Students in the control condition did not discuss their votes with peers, nor did they receive teacher feedback.

This study proceeds with an overview of the literature in Section 3.2. The setup of the experiment and the identification strategy is explained in Section 3.3. Section 3.4 presents the results and Section 3.5 concludes the study and discusses the findings.

## 3.2 Literature

### 3.2.1 Non-explanatory feedback

The aim of providing feedback is to make the learning process visible to students (Hattie & Timperley, 2007). Feedback given does not only provide information about past performances, but also helps students (in the short term) to generate new knowledge and (in the long term) to improve performance (Egelandsdal & Krumsvik, 2019a; Jonsson, 2013). However, feedback does not automatically result in positive effects on student performance. Meta-analysis findings of Bangert-Drowns, Kulik, Kulik and Morgan (1991) revealed that feedback will not affect performance when students are simply told whether their answer is correct or incorrect. Such limited feedback is not explanatory and insufficiently helps students to reflect on course content, to reveal misunderstandings, and to provide information about what is important to learn and what is needed to study further (Bangert-Drowns et al., 1991; Kluger & DeNisi, 1996; Rakoczy, Klieme, Bürgermeister & Harks, 2008). The studies of Chen, Whittinghill and Kadlowec (2010), Oswald, Blake and Santiago (2014) and Yourstone, Krayer and Albaum (2008) all showed that simply giving a polling system (e.g. clicker) to students while not providing explanatory feedback, is not effective in improving learning performance, compared to students who receive explanatory feedback.

### 3.2.2 Teacher feedback

A more typical and commonly used way of using SRS is to show the correct answer and provide teacher feedback (Bachman & Bachman, 2011; Mayer et al., 2009). Before showing the correct answer, voting results are collected and presented visually, for example as a histogram, in front of the class, offering immediate feedback to both students and the teacher. Showing visuals such as histograms helps students to reflect whether they understand the concept and offers them prompts to monitor their level of understanding in relation to their peers (Blasco-Arcas et al., 2013; Sun, 2014). A histogram provides students with feedback in an anonymous way, showing that they are not alone in their thinking, especially when they answered the question incorrectly (Kay & LeSage, 2009; Perez et al., 2010). At the same time, the teacher receives feedback on the students' understanding, which enables him/her to give a specific classroom explanation of the problem when describing the thought processes regarding the correct answer.<sup>6</sup> This feedback informs all students about potential misunderstandings rather than only those students to whom teacher questions are directed (Faber, Luyten & Visscher, 2017). Compared to just showing the given answers and mentioning which one is correct, teacher feedback is meaningful for all students and leads to a better understanding, as teachers have more subject-specific knowledge and use content-specific terms that provide cues on how to solve problems (Chin, 2006; Erdogan & Campbell, 2008; Voerman et al., 2012). Solving problems using a step-by-step demonstration provides clarity to students by identifying gaps in knowledge, helping them to correct misunderstandings and developing an understanding of expectations and standards (Juwah et al., 2004; Nicol & Macfarlane-Dick, 2004). However, even in situations where students answer a question correctly, teacher feedback may contain information on how the question can be addressed more effectively. The studies of Bachman and Bachman (2011), Campbell and Mayer (2009) and Mayer et al. (2009) reported that showing a histogram and hearing teachers providing an explanation for the correct answer positively affects performance on course exams.

---

<sup>6</sup> Feedback can be immediate (directly after answering a question) or delayed (e.g. after a few days or a week). Although immediate feedback is recommended when using SRS (Dabbagh et al., 2019), there are also studies that suggest delayed feedback is more optimal for learning (e.g. Butler & Woodward, 2018; Mullaney, Carpenter, Grotenhuis & Burianek, 2014; Mullet, Butler, Verdin, von Borries & Marsh, 2014). In this study, students receive immediate feedback. It is beyond the scope of this study to investigate the effects of immediate and delayed feedback on learning gains when formatively assessed with SRS.

### 3.2.3 Peer discussions combined with teacher feedback

In a more extensive way of SRS use, feedback is also provided in a peer discussion. When peer discussions and teacher feedback are combined, students first answer a multiple-choice question individually, discuss this question with their peers for a few minutes and then re-vote the question with a new, potentially revised answer based on the conversations with their peers, after which the teacher displays the histogram and reveals and explains the correct answer. This most popular method of SRS use is often called a cooperative approach, based on the work of Mazur (1997). By peer discussions, students get the opportunity to explain and justify their answer. They can try to convince their peers about the correctness of their answer and listen to the views and thought processes of their peers (Crossgrove & Curran, 2008; Egelandstad & Krumsvik, 2019b; Lantz, 2010; Levesque, 2011). This combined approach of peer discussion followed by teacher explanation has been shown to improve student learning (Barth-Cohen et al., 2016; Levesque, 2011; Porter et al., 2013; Smith et al., 2011; Vickrey, Rosploch, Rahmanian, Pilarz & Stains, 2015). During peer discussions, students share and (re)construct knowledge while building on their own understanding or on the idea of a peer. When students clarify their thinking and connect with peers who may have other ideas, they will be exposed to different ways of thinking, and build knowledge that may not have been available before, resulting in a reinforcement of understanding (Barth-Cohen et al., 2016; Chi, De Leeuw, Chiu & LaVancher, 1994; Coleman, 1998; Knight, Wise & Southard, 2013; Tullis & Goldstone, 2020). Previous studies indicated that the processes of self-explanation in peer discussions facilitates students to detect and correct errors in their reasoning (Alevin & Koedinger, 2002; Atkinson, Renkl & Merrill, 2003; Chi et al., 1994; Mathan & Koedinger, 2005). Moreover, recipient students benefit from peer discussions as peers may be more effective explainers than teachers. Peers have a similar background and may be better at finding useful examples or clarifying misunderstandings than a teacher, as peers can describe concepts in a (more) similar language with familiar terms (Caldwell, 2007; Mc Donough & Foote, 2015; Perez et al., 2010; Tullis & Goldstone, 2020). Interactions with peers who may have different ideas or ways of explaining may help students to identify problems that they will not find alone (Knight et al., 2013; Zhonggen, 2017). In a peer discussion pedagogy, where the same question is re-answered, students have to decide (based on possible new ideas) whether to choose the same or a different answer. Several studies have shown that the frequency of correct voting



increases after peer discussions (e.g. Crouch & Mazur, 2001; Mazur, 1997; Egelandstal & Krumsvik, 2017a; Lasry, Charles & Whittaker, 2016; Porter et al., 2011; Smith et al., 2009; Smith et al., 2011; Vickrey et al., 2015; Zingaro & Porter, 2014). Learning gains are found because students in a peer or cooperative condition are more inclined to switch from an incorrect vote to a correct vote than from a correct vote to an incorrect vote (Mazur, 1997; Miller, Schell, Ho, Lukoff & Mazur, 2015; Tullis & Goldstone, 2020). An explanation for these learning gains can be that discussing answers with peers prompts students to clarify their thinking. Through self-explanations, students identify flaws in reasoning and through listening to the views and thought processes of peers, students build knowledge and obtain a deeper understanding of the course content (Chi et al., 1994; Coleman, 1998; Knight et al., 2013). Additionally, students can build on each other's line of reasoning, even when both students initially gave incorrect answers. They may still increase their understanding due to the opportunity of contrasting and comparing their thoughts and ideas (Smith et al., 2009; Versteeg, van Blankenstein, Putter & Steendijk, 2019).

### **3.2.4 Learning from peer discussions**

Another interpretation of 'an increase in learning gains' could be that students simply choose the same answer as their more skilled or more dominating peer during peer discussions; a situation where no learning actually takes place (Nielsen, Hansen-Nygaard & Stav, 2012; Wolfe, 2012). To completely preclude an eventual copying effect of answers, Smith et al. (2009) added a second question, which requires the same conceptual understanding of solutions, but which, like the first question, also needs to be answered individually. They found that students increase their correct answering by 21 percent points when voting this second question in relation to the first question. In a similar study, Egelandstal and Krumsvik (2017) found an improvement of 12 percent points of correctly answering the second question in relation to the first question. The studies of Smith et al. (2009) and Egelandstal and Krumsvik (2017) both showed that students who do not understand concepts at the first individual question, can learn from their peers and transfer acquired knowledge to new similar questions. Nevertheless, it is possible that students discuss incorrect ideas that are not captured in the answer options of the multiple-choice questions, or answer questions that are inconsistent with the ideas they discuss (James & Willoughby, 2011). For these reasons, it seems important for a teacher to combine peer discussions with teacher feedback, as this

combination provides more understanding than the explanation of a peer or teacher alone (Hunsu et al., 2016; Nielsen et al., 2012; Smith et al., 2011; Zingaro & Porter, 2014).

### 3.3 Materials and methods

#### 3.3.1 The design of the experiment

To examine the impact of feedback on students' learning gains in formative assessments with SRS, we use isomorphic questions, in line with the studies of Smith et al. (2011), Porter et al. (2011) and Zingaro and Porter (2014). Accordingly, we set up an intervention with two treatment conditions and one control condition. We evaluate whether teacher feedback (an individual approach) or peer discussion followed by teacher feedback (a cooperative approach) leads to an improvement in student learning in the short term relative to students who receive no feedback. This intervention includes three types of questions:  $Q_1$ ,  $Q_{1PD}$  (*PD* is an abbreviation for Peer Discussion) and  $Q_2$ , as discussed further below.

The intervention took place over a period of 10 weeks. Each week students answered on average four pairs of questions  $Q_1$  and  $Q_2$ . Each concept question  $Q_1$  was followed by a paired isomorphic question  $Q_2$ , which is similar in difficulty or complexity and assesses the same understanding as the concept question  $Q_1$ , but varies in context or numerical values (Smith et al., 2011). Each concept and isomorphic question included four different answer options, of which one is correct. Incorrect answer options include a number of distractors or misconceptions and cannot be eliminated without consideration. In addition, a fifth option "*I don't know*" was included. This latter option should minimize the incidence of blindly guessing when students really do not understand the concept. To get instant feedback from student answers, we used the student response system *Socrative*. This is a free of charge web-based SRS that is available via the website [www.socrative.com](http://www.socrative.com) and can be run on smartphones and laptops. It allows teachers in the *Socrative Teacher module* to easily prepare and manage formative assessments by controlling the flow of questions (Balta & Tzafilkou, 2019; Coca & Sliško, 2017). Students accessed *Socrative* using their smartphones and identified themselves by a six-digit student number in the *Socrative Student module*. As this number was randomly selected by the students themselves, students remained anonymous for the teacher.

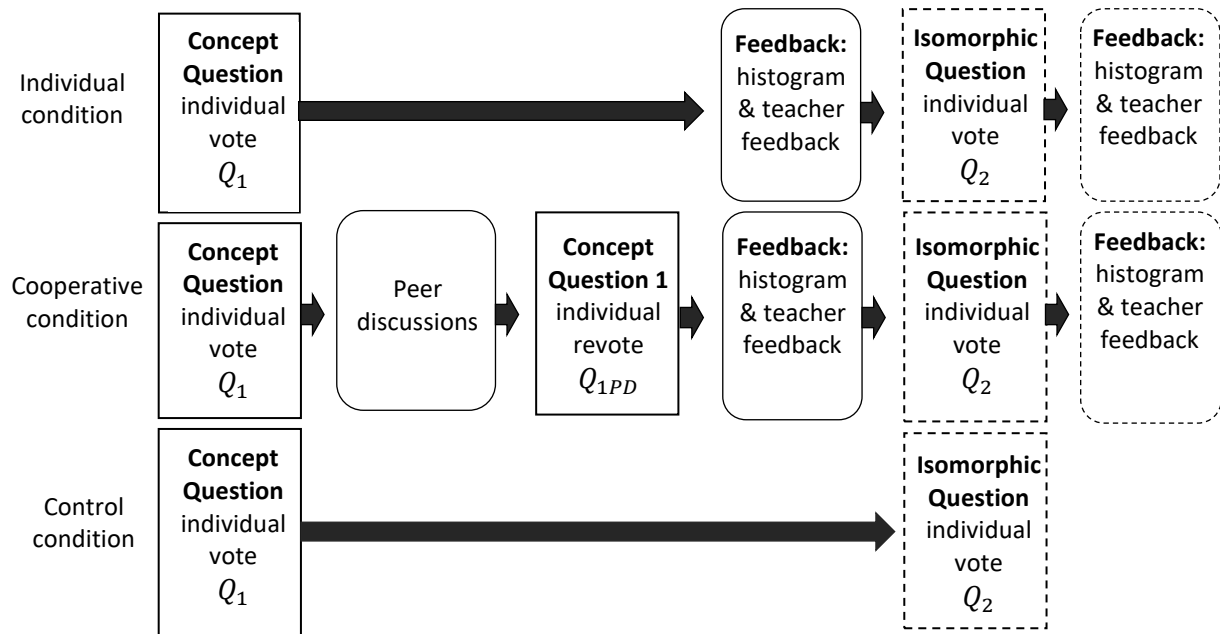
Each multiple-choice question in *Socrative* was displayed on the students' small smartphone screens and on a large screen in front of the classroom. Depending on the difficulty of the question, students were given two to four minutes to consider the question individually, without any interaction between them. All student responses, including null responses, were collected by the *Socrative* software. Although all students in the treated conditions and control condition used *Socrative*, the way in which feedback was provided to the students differs between treatment conditions

Students in the *individual condition* received feedback on concept questions  $Q_1$  of voting results with a histogram. This histogram showed the correct answer and incorrect answers, and a distribution (with percentages) of all anonymous responses. Simultaneously, the teacher explained the correct answer, with feedback aimed at correcting misunderstandings and flaws in logic, focusing on a step-by-step walkthrough of solving the problem (see the upper part of Figure 3.1).

After answering the concept question  $Q_1$  individually, students in the *cooperative condition* did not receive immediate feedback. Instead of showing the histogram and revealing the correct answer, students were encouraged to discuss this question with their peers next to them for several minutes. Students explained and justified their thinking and tried to convince their peer why the answer they chose was correct. In the meantime, the teacher walked around in the classroom, listened to students' reasoning and prompted students to discuss the reasons behind their answers. After discussions, students individually re-voted on the same identical concept question. After submitting a new (and potentially changed) vote  $Q_{1PD}$ , the teacher displayed a histogram of class responses and explained the correct answer, similar to in the individual condition (see the middle part of Figure 3.1).

Students in the *control condition* received no feedback on concept question  $Q_1$ , but only heard the correct answer from the teacher. They saw no histogram, received no problem-solving explanations from their teacher, and did not discuss their solutions with their peers (see the bottom part of Figure 3.1). To avoid confounding of time on task, we agreed with all teachers that the formative assessments in all conditions would take place once a week at the end of a lesson. Each assessment lasted about 30 minutes, regardless of the condition.

Figure 3.1: Description of the experimental design



Next, as a measure of what students individually learned from this process (independent of whether or not they did peer discussions or received teacher feedback), students in all three conditions were asked to individually answer the paired isomorphic question  $Q_2$ . Here, they did not have the opportunity to discuss this question with their peers. After recording the votes, the teachers in the individual and cooperative condition showed the histogram of students’ responses and explained the solution of  $Q_2$ , while the teacher in the control condition only revealed the correct answer without feedback (see the dashed parts of Figure 3.1).

Because students in the control condition did not receive any kind of feedback, we were aware that this could result in potentially penalizing effects. Therefore, we gave all students of all treatment conditions the opportunity to voluntarily view the questions on a website after each class and read the limited written feedback on all answer options. Although we do not have information about whether students made use of this website, we would expect that the natural behavior toward looking at this website is the same in all three conditions, given the complete random assignment of students into classes, and classes into conditions.

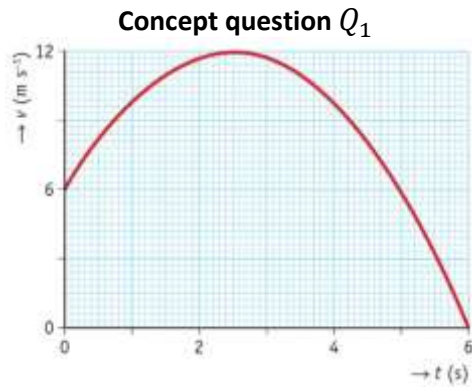
Note that we use the above terms of  $Q_1$ ,  $Q_{1PD}$  and  $Q_2$  throughout the present study.  $Q_1$  refers to an initial individual vote of a concept question,  $Q_{1PD}$  refers to the re-vote of the same

concept question after peer discussion (only for the cooperative condition), and  $Q_2$  refers to an individual vote on a follow-up question. The change in correctness from  $Q_1$  to  $Q_2$  indicates the amount of learning (learning gain) in the short term that occurs during the process of answering paired questions (Porter et al., 2011; Smith et al., 2011; Zingaro & Porter, 2014).

A physics teacher and the researchers developed a database, specifically for this purpose, containing 450 paired multiple-choice questions from which teachers chose questions that related to the current content of their physics course (*mechanics, waves, thermodynamics, electromagnetism or modern physics*). The  $Q_1$  and  $Q_2$  questions pairs originate from multiple-choice questions of standardized physics tests and were selected for higher-order thinking skills (classified at the application and analysis level of Bloom's taxonomy (Bloom, Engelhart, Furst, Hill & Krathwohl, 1956)) that involve deeper conceptual understanding by requiring several thinking steps. In line with the studies of Egelandstad and Krumsvik (2017) and Smith et al. (2011) the order of questions in each pair were randomized to minimize any bias for asking easier  $Q_2$  questions or to eliminate own preferences. Based on this randomization, each questions pair was fixed and categorized as  $Q_1$  and  $Q_2$ . All conditions answered  $Q_1$  first, followed by  $Q_2$ .

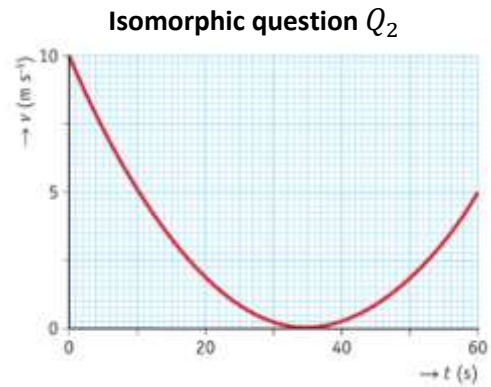
Prior to the intervention, all question pairs used were judged for assessing the same understanding by two independent physics teachers who had experience with writing isomorphic questions. For each question pair, the requirement was that the questions were actually isomorphic and that the same physics principle should be solved. Pairs that were judged by both physics teachers as not testing identical concepts were completely removed from the database by the researchers. The database originally contained 474 question pairs. Both physics teachers independently rated 24 question pairs as not identical. These 24 non-identical question pairs were then completely removed from the database. Regarding the remaining 450 question pairs, we estimated a Cronbach's alpha of 0.81, indicating a high overall internal item consistency. Of these 450 question pairs, both physics teachers rated in total 17 question pairs as different (because one teacher considered the questions as identical, but the other teacher did not). The 17 question pairs were discussed by both teachers and revised such that all question pairs were finally rated as isomorphic. Finally, the database contains 450 question pairs that are considered isomorphic by both physics teachers.

Figure 3.2: Two examples of two paired questions used in this study



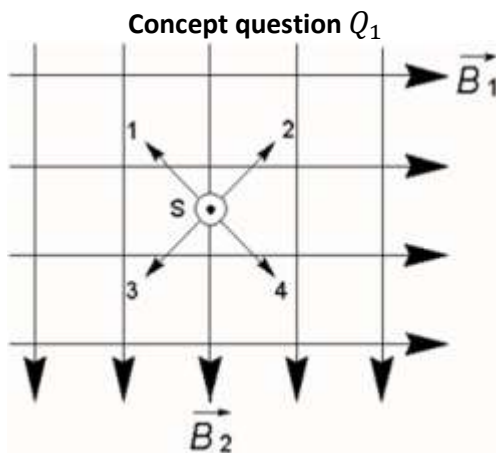
A cheetah moves according to the velocity-time graph shown in the drawing. How far does the cheetah moves at the end of 6 seconds?

- (A)  $1,5 \cdot 10^2$  m
- (B) 17 m
- (C) 52 m**
- (D)  $4,5 \cdot 10^2$  m
- (E) I don't know



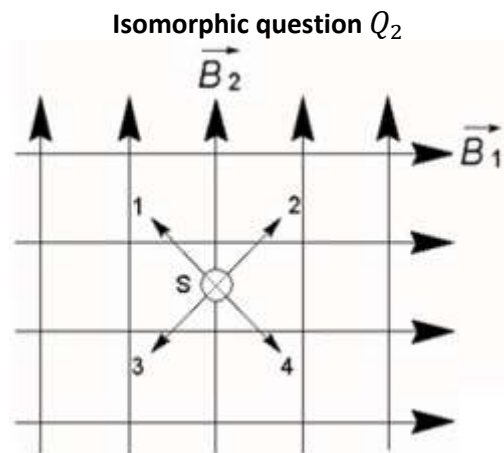
A motorcycle travels according to the velocity-time graph shown in the drawing. How far does the motorcycle travels at the end of 60 seconds?

- (A)  $1,9 \cdot 10^2$  m
- (B)  $1,6 \cdot 10^2$  m**
- (C)  $4,5 \cdot 10^2$  m
- (D) 66 m
- (E) I don't know



A long straight current wire is located in two equally strong homogeneous magnetic fields  $B_1$  and  $B_2$ . The current in the wire is directed out of the paper. Which arrow shows the direction of the resulting magnetic force  $F_L$  on the wire in S?

- (A) arrow 1
- (B) arrow 2**
- (C) arrow 3
- (D) arrow 4
- (E) I don't know



A long straight current wire is located in two equally strong homogeneous magnetic fields  $B_1$  and  $B_2$ . The current in the wire is directed in the paper. Which arrow shows the direction of the resulting magnetic force  $F_L$  on the wire in S?

- (A) arrow 1
- (B) arrow 2
- (C) arrow 3
- (D) arrow 4**
- (E) I don't know

Each week, when preparing a formative assessment, teachers were completely free to select four paired sets ( $Q_1$  and  $Q_2$ ) of multiple-choice questions that matched the content of their lessons. Each pair of questions was selected as a fixed pair in this process. See Figure 3.2 for two examples of paired  $Q_1$  and  $Q_2$  questions. The correct answers in Figure 3.2 are in bold typeface.

### 3.3.2 Data

Our database contained 12,253 responses on question pairs. In order to control for differences between the students that answer the questions, we matched the students' demographic information (gender, age, education level and physics test score of previous year) to each answered question pair by means of the unique six-digit student number. A closer look at the data showed that not all the answered question pairs met our requirements, as the data showed that in 585 occasions of responses a student in the cooperative condition did not mention the student number of the peer.<sup>7</sup> Reasons for this could be that they forgot to fill in the number of the peer or that they were not sitting next to another student who acted as a peer, which are actually two very different situations with respect to getting feedback from a peer or not. Because it is not clear whether or not these questions the  $Q_1$  vote was discussed with a peer, we excluded the 585 responses from our database, without violating the assumptions underlying the RCT.<sup>8</sup> Furthermore, 67 responses of 7 questions pairs that were answered 100 % correctly on average in a condition were also excluded from the database, as there were no learning gains possible.<sup>9</sup> This concerned 14 responses on 1 question pair from students in the control condition, 35 responses on 3 question pairs from students in the

---

<sup>7</sup> Answered question pairs ( $Q_1$ ,  $Q_{1PD}$ ,  $Q_2$ ) in the cooperative condition were linked between pairs of students that acted as each other's peers, which is possible since students in this condition identified themselves in *Socratic* by their student number followed by the student number of their peer.

<sup>8</sup> To check whether students in the cooperative condition who answered question pairs without peer discussions are a specific group with respect to one or more of the background characteristics, we carry out  $t$ -tests between these students and those students who answered all question pairs with peer discussions. We do not find significant differences in student characteristics between groups. See Table 3.1 for the  $t$ -tests;  $p > 0.05$  in all tests. We also conducted additional  $t$ -tests (not shown in Table 3.1) between a group of students ( $N = 123$ ) who answered more than 10 % of all question pairs without peer discussions with a group of students ( $N = 124$ ) who answered less than 10 % of all question pairs without peer discussions. The chosen value of 10 % is based on the median. Similar to the results in Table 3.1, we find no significant differences between student characteristics between the two groups;  $p > 0.05$  in all tests.

<sup>9</sup> The exclusion of these 7 question pairs is a result of how learning gains are defined and measured. This is explained in Section 3.3.5.

individual condition and 18 responses on 3 question pairs from students in the cooperative condition.

The descriptive statistics of the remaining number of answered question pairs are summarized in Table 3.2. The final database contains in total 11,601 responses on question pairs and includes 1,942 responses on 89 question pairs in the control condition, 4,289 responses on 155 question pairs in the individual condition and 5,370 responses on 195 question pairs in the cooperative condition. A total of 439 question pairs have been answered in one or more of our treatment conditions. Of these 439 question pairs, 99 question pairs (with a total of 2,709 responses) are answered in one or more classes in all three treatment conditions (not visible from Table 3.2).

Table 3.1: T-tests between students in the cooperative condition

	<i>Students who answered one or more question pairs without peer discussions (N = 141)</i>		<i>Students who answered all question pairs with peer discussions (N = 106)</i>		<i>t</i>	<i>p</i>
	<i>Mean</i>	<i>Std. Dev.</i>	<i>Mean</i>	<i>Std. Dev.</i>		
Gender <sup>A</sup>	0.51	0.50	0.48	0.49	0.46	0.65
Age	15.87	1.04	15.72	1.06	1.10	0.27
Education level <sup>B</sup>	0.74	0.44	0.67	0.47	1.16	0.25
Physics test score	6.62	1.05	6.61	1.02	0.10	0.92

\*\*\* = significant at the 1 % level; \*\* = significant at the 5 % level; \* = significant at the 10 % level.

<sup>A</sup> boy = 0, girl = 1

<sup>B</sup> general secondary education = 0, pre-university education = 1

- T-tests between students in the cooperative condition who answered one or more question pairs without peer discussions with students in the cooperative condition who answered all question pairs with peer discussions.

- 141 students in the cooperative condition answered one or more question pairs without peer discussions. These students answered a total of 1 to 7 question pairs without a peer.

Table 3.2: Descriptive statistics of answered question pairs

	<i>N<sub>s</sub></i> <i>(number of students in the condition)</i>	<i>N<sub>q</sub></i> <i>(asked number of question pairs)</i>	<i>n</i> <i>(responses on question pairs of all students)</i>	<i>Min.</i> <i>(answered question pairs of a student)</i>	<i>Max.</i> <i>(answered question pairs of a student)</i>
Control condition	111	89	1,942	8	40
Individual condition	169	155	4,289	8	40
Cooperative condition	247	195	5,370	6	40



### 3.3.3 Descriptive statistics

Based on a survey, about 48 % of the 527 students are girls. None of the students indicated a gender other than *boy* or *girl*. At the start of the experiment, the students are on average 15.75 years old ( $SD = 0.94$ ). Seventy percent of them are enrolled in the pre-university education track (six years; the highest track in Dutch secondary education), the remainder of the students are enrolled in the general secondary education (five years; the middle track in Dutch secondary education). The average score on physics tests in the previous year is 6.61 ( $SD = 0.96$ ), where grades range between 1 and 10 (10 is outstanding and 5.5 is barely sufficient for passing).

The quality of the randomization was examined using analysis of variance (ANOVA) tests. Table 3.3 presents the comparison of students' pre-treatment variables of the control condition and the two treatment conditions. The ANOVA tests show that students in the control conditions are, on average, similar to students of the treated conditions. In Sections 3.4.1 and 3.4.2, we will control for all these variables in regressions to further reduce potential omitted variable bias.

Table 3.3: Comparison between untreated and treated conditions

	<i>Control condition</i> ( <i>N</i> = 111)		<i>Individual condition</i> ( <i>N</i> = 169)		<i>Cooperative condition</i> ( <i>N</i> = 247)		<i>F</i>	<i>p</i>
	<i>Mean</i>	<i>Std. Dev.</i>	<i>Mean</i>	<i>Std. Dev.</i>	<i>Mean</i>	<i>Std. Dev.</i>		
Gender <sup>A</sup>	0.52	0.50	0.43	0.50	0.50	0.50	1.55	0.21
Age	15.61	0.67	15.78	0.93	15.80	1.05	1.61	0.20
Education level <sup>B</sup>	0.68	0.47	0.72	0.45	0.71	0.46	0.28	0.75
Physics test score	6.67	1.03	6.56	0.76	6.62	1.04	0.43	0.65

\*\*\* = significant at the 1 % level; \*\* = significant at the 5 % level; \* = significant at the 10 % level.

<sup>A</sup> boy = 0, girl = 1

<sup>B</sup> general secondary education = 0, pre-university education = 1

### 3.3.4 Participant and randomization

The study was conducted in six secondary schools in the southern part of the Netherlands. All schools are located in an area, outside the highly urbanized, central region of the Netherlands and are representative of the average secondary school in the Netherlands.<sup>10</sup> In total, 629 students in 30 classes, belonging to the 10<sup>th</sup> to 12<sup>th</sup> grade (upper secondary education) of general higher secondary education and pre-university education, are part of the experiment. The intervention lasted 10 weeks. One week before the start of the experiment, students voluntarily filled in a survey in the presence of the teacher. This survey gathered demographic information (their self-chosen six-digit student number they use to log in *Socrative*, gender [students were given three response options: *boy*, *girl*, and *other*], age, education level and physics test score of previous year).

We excluded 92 students from the analysis because they hardly participated in the weekly assessments ( $N = 57$ ), left school or class during the duration of the experiment ( $N = 6$ ) or did not fill out the survey of the experiment ( $N = 29$ ). On top of that, another 10 students are excluded because they logged each assessment with pseudonyms (and not with their self-selected six-digit student number) into *Socrative*, so that the data from different assessments were not traceable to the students. The final sample includes 527 students who are nested within 30 classes and taught by 12 teachers. Of these 30 classes, nine classes ( $N = 169$ ) are in the individual condition, fifteen classes ( $N = 247$ ) are in the cooperative condition and six classes ( $N = 111$ ) are in the control condition.

Well before the start of the intervention, seven schools were selected, based on previous contacts of the authors, and contacted to determine their potential interest to participate in this study. The principals were reassured that the study would not interfere with other issues at school. Six schools consented to participate and agreed with the treatment conditions of

---

<sup>10</sup> Compared with the average Dutch secondary school ( $N = 648$ ), the averages of the participating schools ( $N = 6$ ) in this study are: Total students:  $M = 1,658$  (national average:  $M = 1,504$ ,  $SD = 1,142$ ); Teachers employed:  $M = 132$  (national average:  $M = 114$ ,  $SD = 92$ ); Teachers age:  $M = 44.9$  (national average:  $M = 42.0$ ,  $SD = 10.0$ ); Graduation percentage:  $M = 87.9\%$  (national average:  $M = 89.3\%$ ,  $SD = 5.0$ ); Exam grade:  $M = 6.61$  (on a scale from 1 to 10) (national average  $M = 6.50$ ,  $SD = 0.41$ ). All statistics are within half a standard deviation of the average of all variables. This indicates that the participating schools are representative of the average Dutch secondary school. (The data are from 2018, the same year the intervention took place, and are obtained from the governmental website: [https://www.duo.nl/open\\_onderwijsdata/databestanden/vo/](https://www.duo.nl/open_onderwijsdata/databestanden/vo/).)

the different groups. The principals of each school nominated the teachers and classes to participate in this study. To ensure that the intervention and the use of the SRS would be used correctly, all teachers participated in a one-hour training session, which included an explanation of the motivation for the intervention, the requirements pertaining to the feedback under which the treatments had to occur and a training in how to use *Socratic* in the classroom. The teachers also received the database with paired multiple-choice questions for preparing the formative assessments. We also explained to the principals of the schools the importance of randomization of the participants, and suggested randomizing students over classes before the timetables of the courses were made, to prevent timetabling issues interfering with the possibility of randomization. Therefore, the students at each school were first randomly assigned by scheduling software to one of the physics classes. The software randomly divided students, based on their courses and education level. Next, the researchers randomly assigned classes within schools to the individual condition, the cooperative condition or the control condition. As students are randomly assigned to classes, and classes are randomly assigned to teachers, in fact, students are also randomly assigned to questions pairs. By randomizing classes in this way, we tried to minimize possible teacher effects, as selection into the treated and untreated conditions did not depend on the teacher. By including six schools in the experiment, and by randomizing within schools, we minimize possible school effects and increase the validity of our study.

### **3.3.5 Outcome variable**

Students' learning gains were assessed using the individual student responses in the weekly formative assessments. Each assessment contained on average four pairs of questions. For each pair, we used the aforementioned notation  $Q_1$ ,  $Q_{1PD}$ , and  $Q_2$ , with the first vote (and second vote in the cooperative condition) taken on the first concept question and the second vote (and third vote in case of the cooperative condition) taken on the second isomorphic question. A score of one (1) was given for each question answered correctly and a score of zero (0) was given for each question answered incorrectly.

To estimate the effects on learning gains of teacher feedback, or a combination of peer discussions and teacher feedback, we follow the studies done by Murnane and Willett (2010). In particular, we calculate the average normalized learning gain on each of the question pairs

and for every treatment and control condition separately. In total 89 question pairs ( $p$ ) were solved by 111 students  $i$  in the control condition, 155 question pairs by 169 students  $i$  in the individual (treatment) condition and 195 question pairs by 247 students  $i$  in the cooperative (treatment) condition (see Table 3.2). Because students are randomly assigned to classes, and classes are randomly assigned to teachers, the students are also randomly assigned to question pairs. Considering the calculations of the average normalized learning gains at the level of the question pair, the students are nested in question pairs  $p(i)$ , and we then may write:

$$LG_{p(i)} = \frac{E[Q_{2i}=1 | Q_{1i}=0]}{E[Q_1=0]} \quad (3.1)$$

Equation (3.1) shows that the average normalized learning gains are expressed as the likelihood that the isomorphic question  $Q_2$  got answered correctly by a student  $i$ , in the event that this student  $i$  made a mistake on question  $Q_1$ , and conditional on the fact that student  $i$  had to answer the question pair  $Q_{1,2}$ . By including the denominator in Equation (3.1) into the calculation of the learning gains, we account for the fact that not all students answered the same question pairs in our database. Different classes may have used different question pairs in accordance with the lessons covered. For example, some classes deal with more difficult learning material than other classes and, consequently, some classes may get different questions pairs than other classes.

Owing to the randomization procedure, and also to the way we calculate the normalized learning gains, differences in the learning gains between the two treatments and control conditions can only be ascribed to the treatment and not to, for example, differences in background characteristics of students answering those question pairs (Murnane & Willett, 2010). Furthermore, since the growth in learning between  $Q_1$  and  $Q_2$  is measured for the same students, we implicitly can control for time invariant characteristics of those students in Equation (3.1). This allows us to further reduce potential omitted variable bias. These claims are tested in several ways in Section 3.4.1 and Section 3.4.2, and find that our results are robust to different specifications.

In summary, we have a database with 11,601 student responses on 439 question pairs. From this database we calculate the mean normalized learning gains according to Equation (3.1), having a database then with 439 question pairs, complemented by (1) the average normalized learning gains of each condition separately on each of the question pairs, and (2) the mean scores of pre-treatment variables  $X_{zp(i)}$  of students who answered the corresponding question pairs. The pre-treatment characteristics of students will mainly be used in the regression analysis to check for robustness of the results.

As such, we estimate the impact on learning gains (Model 1) of teacher feedback with or without peer discussions by looking at the differences in learning gains between the two treatment conditions and the control condition in a multivariate regression:

$$LG_{p(i)} = \alpha_0 + \beta_1 T_{1p(i)} + \beta_2 T_{2p(i)} + \delta + \varepsilon_{p(i)} \quad (3.2)$$

$T_{p(i)}$  corresponds to the treatment conditions, namely, the individual condition  $T_1$ , and the cooperative condition  $T_2$ . The parameter  $\delta$  indicates that we control for time invariant effects, such as background characteristics of a student, or the level of difficulty of the question pairs by the way the average normalized learning gains are calculated in Equation (3.1).  $\varepsilon_{p(i)}$  is the usual standard error of the regression.

To improve the precision of the estimates  $\beta_1$  and  $\beta_2$ , we add a vector of covariates  $X_{zp(i)}$  to the multivariate regressions. These covariates are all measured before the experiment took place, often referred to as pre-treatment characteristics, as to avoid any interferences with the treatments. The estimate of the treatment effect should be comparable with and without these covariates, owing to the randomization procedure, and, consequently, Equation (3.3) is mainly used as a robustness analysis.

$$LG_{p(i)} = \alpha_0 + \beta_1 T_{1p(i)} + \beta_2 T_{2p(i)} + \sum_{z=1}^Z \gamma_z X_{zp(i)} + \delta + \varepsilon_{p(i)} \quad (3.3)$$

We have a set of covariates that we add to the model stepwise. We add the average female population and the average age of students that solved a question pair to the regression in Model 2. Likewise, we also add the average educational level and the average physics test score of previous year to the regression in Model 3. Even though these covariates are not significantly different between the treated and untreated conditions, they still contribute to

the outcome  $LG_{p(i)}$  and make the models better performing (Bloom, Richburg-Hayes & Black, 2007; Raudenbush, 1997). To allow direct comparison of the findings, the scores of all variables included in the analysis are converted to standardized (z-) scores, except the dichotomous student variables gender (girl = 1) and education level (pre-university education = 1). In Model 4A, we estimate fixed effects by including time dummies in the analysis, which are dummies indicating in which week the question pair was answered.<sup>11</sup> By doing so, we control for all unobserved student characteristics that are invariant or ‘fixed’ over time (Balta, Michinov, Balyimez & Ayaz, 2017). This allows us to double check our reasoning that time invariant effects are already controlled for by the way we measure learning gains, because in that case week effects should not significantly alter the estimated coefficients of the treatment conditions. In Model 4B, we estimate teacher fixed effects to control for unmeasured time-invariant teacher characteristics.

We must note that the database of 11,601 responses to pairs of questions has a hierarchical structure, as students are nested in classes and classes are nested in schools. Therefore, it is not inconceivable to use multilevel regression analyses. However, in this study we use multivariate regressions. In Appendix 3.1, we explain why multivariate regressions are preferred over multilevel regressions.

## 3.4 Results

### 3.4.1 Results learning gain

Table 3.4 summarizes the results of the effect of feedback strategies with SRS on learning gains (the average normalized learning gains  $LG_{p(i)}$  between  $Q_1$  and  $Q_2$ , given the difficulty of the question pairs) in the short term. Model 1 is a basic model that only includes the treatment status of the students. Compared to the control condition, a significant positive treatment effect is observed in the individual condition ( $\hat{\beta}_1 = 0.25, p < 0.01$ ) and in the cooperative condition ( $\hat{\beta}_2 = 0.34, p < 0.01$ ). This implies that teacher feedback, whether or not combined

---

<sup>11</sup> Formative assessments were conducted over ten weeks. It is possible that time effects play a role in explaining our measured effects. Students could perform better/worse in the last weeks of the formative assessments than at the beginning of the assessments. Reasons for this could be diverse (more/less motivation, more/less anxiety, more/less familiarity, etc.). In order to examine whether any time effects play a role, we add time dummies to the analysis.

with peer discussions, positively affects learning outcomes compared with the control condition which do not receive any kind of feedback. In Model 2, we increase the precision of these estimates by including the covariates gender and age to the analysis (we remind that gender is the average female population and that age refers to the average age of the students who answered the question pairs). The effects of the treatments slightly decrease, but the significance is retained with  $\hat{\beta}_1 = 0.21$  ( $p < 0.01$ ) and  $\hat{\beta}_2 = 0.29$  ( $p < 0.01$ ), respectively. Furthermore, gender ( $\hat{\gamma}_1 = -0.31$ ,  $p < 0.01$ ) and age ( $\hat{\gamma}_2 = 0.051$ ,  $p < 0.05$ ) are significant predictors of learning gain. This means that girls score 31 percent lower on learning gains compared to boys, while learning gains increase by 5.1 % as age increases by 1 standard deviation.

Table 3.4: Multivariate regression analyses predicting learning gains

	Model 1 <i>Learning gains<sup>A</sup></i>	Model 2 <i>Learning gains<sup>A</sup></i>	Model 3 <i>Learning gains<sup>A</sup></i>	Model 4A <i>Learning gains<sup>A</sup></i>	Model 4B <i>Learning gains<sup>A</sup></i>
( $\beta_1$ ) Individual condition	0.25*** (0.038)	0.21*** (0.042)	0.17*** (0.043)	0.17*** (0.043)	0.19*** (0.038)
( $\beta_2$ ) Cooperative condition	0.34*** (0.036)	0.29*** (0.040)	0.29*** (0.039)	0.29*** (0.038)	0.28*** (0.034)
( $\gamma_1$ ) Gender (boy = 0 & girl = 1)	--	- 0.31*** (0.11)	- 0.47*** (0.12)	- 0.43*** (0.12)	- 0.34*** (0.086)
( $\gamma_2$ ) Age	--	0.051** (0.020)	- 0.0018 (0.026)	- 0.0086 (0.026)	- 0.0062 (0.020)
( $\gamma_3$ ) Education level	--	--	0.12*** (0.038)	0.13*** (0.038)	0.11*** (0.030)
( $\gamma_4$ ) Physics test score prev. year	--	--	- 0.13** (0.055)	- 0.14** (0.055)	- 0.069 (0.052)
Fixed effects	no	no	no	yes time	yes teacher
Constant	0.31*** (0.031)	0.49*** (0.072)	0.48*** (0.071)	0.46*** (0.086)	0.43*** (0.052)
Observations	439	439	439	439	439
R <sup>2</sup>	0.18	0.21	0.26	0.26	0.26
F	43.13 (2, 429)	27.56 (4, 427)	24.42 (6, 425)	11.16 (15, 416)	11.13 (17, 414)

\*\*\* = significant at the 1 % level; \*\* = significant at the 5 % level; \* = significant at the 10 % level.

<sup>A</sup>The control condition is the reference category of the treatment conditions.

Adding the covariates 'education level' and 'physics test score previous year' (which are again average values of pre-treatment characteristics of students who answered the paired questions) to the analysis show that the effect of the treatment in Model 3 decreases for the individual condition ( $\hat{\beta}_1 = 0.17, p < 0.01$ ) and remains unchanged for the cooperative condition ( $\hat{\beta}_2 = 0.29, p < 0.01$ ). Furthermore, gender ( $\hat{\gamma}_1 = -0.47, p < 0.01$ ), education level ( $\hat{\gamma}_3 = 0.12, p < 0.01$ ) and physics test score previous year ( $\hat{\gamma}_4 = -0.13, p < 0.05$ ) are significant predictors of learning gains. This means that students with a pre-university education (six years) experience 12 % more learning gains than students with a general higher secondary education (five years), which is a significant difference. Furthermore, learning gains decrease by 13 % as students increase their physics score in previous year by 1 standard deviation. By adding the variable education level in Model 3, we notice that the variable age loses its significance. The reason for this is that education level significantly correlates with age ( $\rho = 0.32, p < 0.01$ ), since a part of the students who are in pre-university education (six years) are as well older than students who are in general higher secondary education (five years).

To see whether we control for all unobserved student characteristics that are constant over time, as hypothesized in Section 3.3.5, we control in Model 4A for time dummies. The analysis in Model 4A shows that the estimates of  $\hat{\beta}$  remain constant as week dummies are included. As an additional analysis, we also control for teacher fixed effects in Model 4B. This does not change our conclusions, with estimates of  $\hat{\beta}_1 = 0.19 (p < 0.01)$  and  $\hat{\beta}_2 = 0.28 (p < 0.01)$ . From this we can conclude that results are not driven by differences in teacher characteristics across classrooms and schools.

### 3.4.2 Robustness analyses of learning gain

To check further robustness of our results, we perform two additional analyses which are shown in Table 3.5. The first robustness check that we perform is an analysis for which we do not use all answered question pairs of our data, but only those pairs that are answered in all three conditions. It concerns in total 33 of the same question pairs in both the control condition, the individual condition and the cooperative condition. This limited number of question pairs is only 22.5 % of all questions pairs in total, but the results of the first analysis produce similar results for the treated conditions to those in Table 3.4. We have argued in Section 3.3.5 that by the way we have measured learning gains, student ability and the level



Table 3.5: Robustness analyses with outcome learning gain

Robustness analyses	<i>Learning gains<sup>A</sup></i> <i>Only paired questions asked in all three conditions</i>	<i>Learning gains<sup>A</sup></i> <i>Only paired questions answered with a <math>Q_1</math> correctness of &lt; 80 %</i>
$(\beta_1)$ Individual condition	0.13** (0.078)	0.15*** (0.036)
$(\beta_2)$ Cooperative condition	0.33*** (0.081)	0.25*** (0.031)
$(\gamma_1)$ Gender (boy = 0 & girl = 1)	- 0.50* (0.30)	- 0.40*** (0.11)
$(\gamma_2)$ Age	- 0.041 (0.060)	- 0.029 (0.023)
$(\gamma_3)$ Education level	0.085 (0.089)	0.12*** (0.035)
$(\gamma_4)$ Physics test score prev. year	- 0.22 (0.15)	- 0.12** (0.047)
Fixed effects <sup>B</sup>	yes	yes
Constant	0.51 (0.16)	0.38*** (0.062)
Observations	99	332
$R^2$	0.26	0.28
$F$	5.96 (6, 89)	23.79 (6, 318)

\*\*\* = significant at the 1 % level; \*\* = significant at the 5 % level; \* = significant at the 10 % level.

<sup>A</sup> The control condition is the reference category of the treatment conditions.

<sup>B</sup> Fixed effects are measured by including week dummies. There are ten week dummies in total.

of difficulty of question pairs no longer play a role in explaining the results of learning gains  $LG_{p(i)}$ . By randomly assigning students to classes, and, hereby, also to question pairs, we are allowed to include question pairs that are not asked in all treatment conditions (as shown in Table 3.4).

The second robustness check in Table 3.5 is an analysis that is done to check that the effects we find are not due to the fact that we have questions that are too easy in our data. As noted by Crouch and Mazur (2001), Smith et al. (2011) and Zingaro and Porter (2014), isomorphic questions that are too easy are insufficiently challenging and leave little opportunity for gains in learning. Although there is no standard cutoff for 'too easy' (Zingaro & Porter, 2014), Smith et al. (2011) and Zingaro and Porter (2014) dropped questions where the individual  $Q_1$  vote is

more than 80 % correct. If we also choose to do this, we exclude 107 question pairs out of 439 (10 question pairs from students in the control condition, 36 question pairs from students in the individual condition and 61 question pairs from students in the cooperative condition). The results of the second robustness check with 'sufficiently challenging' (Zingaro & Porter, 2014) question pairs show effects similar to those presented in Table 3.4.

All in all, we show that the results of learning gains  $LG_{p(i)}$  are robust to several model specifications and covariates. Therefore, we can conclude that teacher feedback or a combination of peer discussions and teacher feedback have a positive significant effect on learning gains in the short term.

### 3.4.3 Further evidence on potential learnings

In this section, we aim to provide insights into the contribution of peer discussions to the comprehension of concepts. Figure 3.3 plots flowcharts of students' response patterns of paired questions in the untreated and treated conditions. The flowcharts of the control condition and individual condition trace response patterns of concept questions  $Q_1$  and paired isomorphic questions  $Q_2$ . The flowchart of the cooperative condition is expanded with response patterns of questions  $Q_{1PD}$  (the re-votes of  $Q_1$  after peer discussion). To make fair comparisons between results, we only included the 99 question pairs that are asked in all treatment conditions (33 question pairs in each separate condition).<sup>12</sup> By doing so, we only compare responses on question pairs of the same learning material and the same level of understanding.

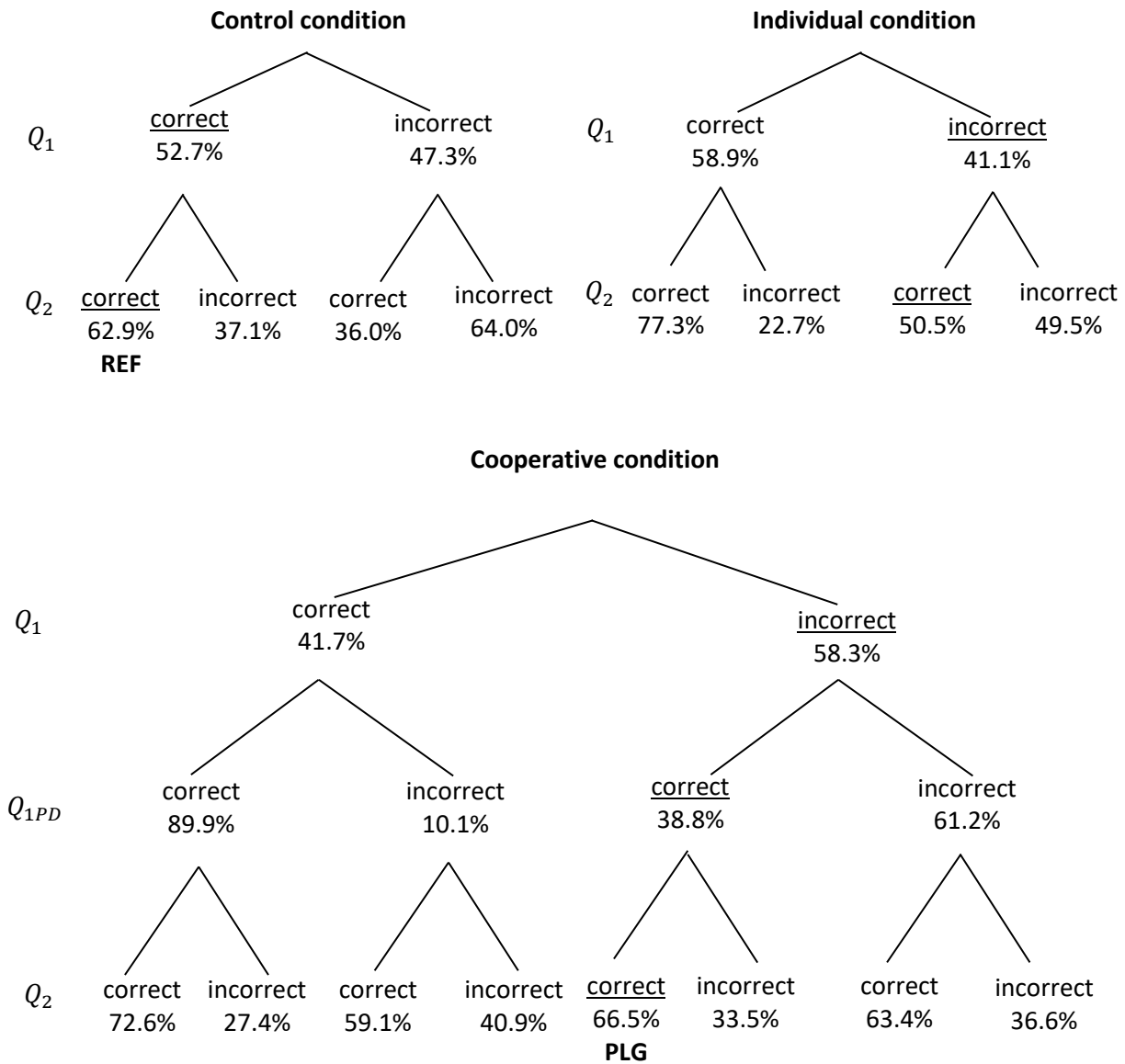
The top two branches of all flowcharts correspond to the percentages of  $Q_1$  questions that are answered correctly (left branch) and incorrectly (right branch). The bottom branches correspond to the percentages of split subgroups of follow-up  $Q_2$  questions that are answered correctly and incorrectly. The flowchart of the cooperative condition includes a middle layer with branches of percentages of  $Q_{1PD}$  questions that are answered correctly (left branches) and incorrectly (right branches). The percentages in Figure 3.3 are relative and compare the averages across all responses. For example, in the cooperative condition, 89.9 % of the 41.7 %

---

<sup>12</sup> The 99 question pairs included a total of 2,709 student responses; 716 student responses in the control condition, 950 student responses in the individual condition and 1,043 student responses in the cooperative condition.

correctly answered  $Q_1$  questions are also answered correctly after peer discussions. Of this subgroup, another 72.6 % of the  $Q_2$  questions are also answered correctly.

Figure 3.3: Flowcharts of untreated and treated conditions



In the light of the analysis of the contribution of the peer discussion to the comprehension of concepts, a group worthy of further study is the group of students in the cooperative condition who answer  $Q_1$  incorrectly and  $Q_{1PD}$  correctly. Porter et al. (2011) defined this group as the ‘potential learner group’ (PLG), as they can have the potential to learn from peer discussions. That is, they learn from their peers and master the concepts of  $Q_1$  if they are able to transfer acquired knowledge to the new isomorphic question  $Q_2$ . The flowchart of the cooperative

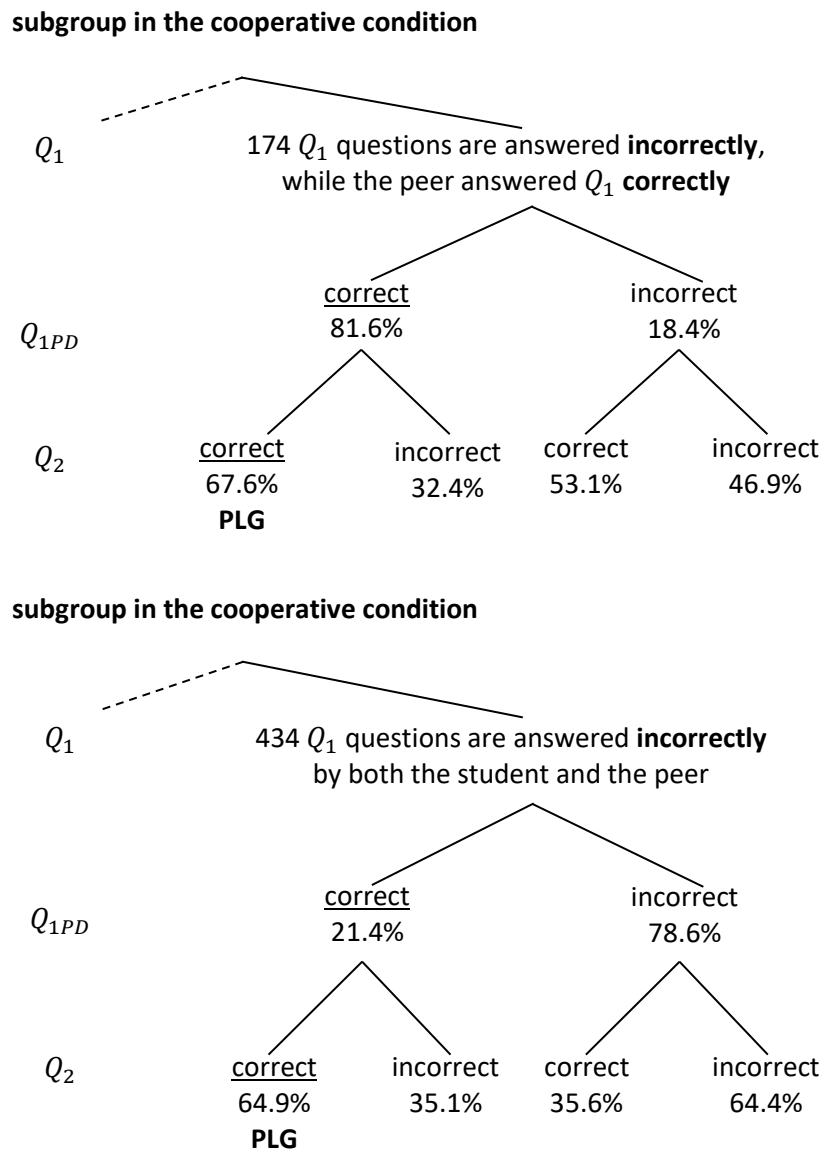
condition in our study shows that 66.5 % of the potential learner group answer  $Q_2$  correctly (Figure 3.3). That is, over two-thirds of these students learn from peer discussions, combined with teacher feedback on  $Q_{1PD}$ . The extent to which this potential learner group learns can be determined by comparing them to a subgroup of students who already master the concepts of  $Q_1$  from the start. By doing this, we select those students in the control condition who answered  $Q_1$  correctly (Porter et al., 2011; Zingaro & Porter, 2014). These students belong to our reference group (REF), as they are expected to answer  $Q_2$  correctly, without receiving feedback on the previous question  $Q_1$ . However, the flowchart in the control condition shows that not all students in this reference group answer  $Q_2$  correctly; only 62.9 % of the students who answer  $Q_1$  correctly do so (Figure 3.3). It is recommended by Porter et al. (2011) and Zingaro and Porter (2014) to normalize the expectations of the potential learner group to answer  $Q_2$  correctly based on this number. That is, as not all questions  $Q_2$  are answered correctly in the reference group by the students that answered  $Q_1$  correctly, we should reduce our expectations for the potential learner group in the cooperative condition.<sup>13</sup> Doing so, we find that the potential learner group in the cooperative condition is 106 % (= 66.5/62.9) as likely to answer  $Q_2$  correct as the reference group in the control condition. This implies that proportionally more potential learners in the cooperative condition answer  $Q_2$  correct than students in the control condition who understood the concept from the beginning.

It is also possible that students in the potential learner group (PLG) may answer  $Q_2$  correctly due to teacher feedback on  $Q_{1PD}$ , and not as a result of peer discussions (because they simply choose the same correct answer from their more skilled peer, without understanding the concept). In order to find out whether the potential learner group in the cooperative condition comprehends concepts from peer discussions, we examine a subgroup of students who answer  $Q_1$  incorrectly, but in which the peers answer  $Q_1$  correctly (the upper part of Figure 3.4), and another subgroup of students in which  $Q_1$  is answered incorrectly by both the student and the peer (the bottom part of Figure 3.4).

---

<sup>13</sup> Porter et al. (2011) and Zingaro and Porter (2014) stated that the percentage of the reference group that correctly answer  $Q_2$  is a measure of the maximum of available learning for the potential learner group (PLG). They calculated the weight of performance of the potential learner group (PLG) on the weight of performance of the reference group (REF) with:  $\frac{\% \text{ correct PLG}}{\% \text{ correct REF}}$ . Porter et al. (2011) and Zingaro and Porter (2014) both indicated that this measure of maximum available learning is more representative of the value of discussion than the raw percentages.

Figure 3.4: Flowcharts of two subgroups in the cooperative condition



We compare both subgroups to a subgroup of students of the individual condition who answer  $Q_1$  incorrectly and  $Q_2$  correctly. This subgroup in the individual condition does not understand the concept of  $Q_1$  (as do students in the potential learner groups in Figure 3.4), but learn from teacher feedback (as well as the students in the potential learner groups). The flowchart of the students in the potential learner group who answered  $Q_1$  incorrectly while the peer answered  $Q_1$  correctly (upper part in Figure 3.4) shows that 81.6 % of the initially incorrect answers  $Q_1$  are changed in correct answers  $Q_{1PD}$ . Of this subgroup, another 67.6 % of the  $Q_2$  questions are also answered correctly.

Another student response pattern is observed when  $Q_1$  is answered incorrectly by both the student and the peer (bottom part in Figure 3.4); 21.4 % of the answers are changed into a correct  $Q_{1PD}$ . However, of this subgroup, another 64.9 % of the  $Q_2$  questions are answered correctly. When we compare the flowcharts of two potential learner groups (PLGs) in Figure 3.4 with the flowchart of the individual condition (Figure 3.3), we observe that the 50.5 % incorrect answers  $Q_1$  that are changed into correct answers  $Q_2$  in the individual condition is less than the 67.6 % and 64.9 % of the potential learner groups (Figure 3.4). Based on these numbers, we hypothesize that the potential learners in the cooperative condition could learn from the peer discussions, and not solely from the teacher feedback.

### 3.5 Conclusion and discussion

The purpose of the present study was to analyze whether SRS supported assessment activities are effective in increasing learning gains in the short term. Therefore, we conducted a randomized experimental trial with sufficient statistical power in upper secondary education. Our first research question was in what extent SRS supported assessment activities enhance students' learning gains. Therefore, we evaluated whether receiving teacher feedback on concept questions (an individual approach), or peer discussions followed by teacher feedback (a cooperative approach) led to an improvement of comprehension when answering a new isomorphic question relative to students who do not discuss their votes with peers, or do not receive teacher feedback (the control condition). The findings show that there are significant positive effects of teacher feedback, whether or not combined with peer discussions, on learning gains compared to students who receive no feedback. The results imply a Cohen's  $d$  effect size of 0.83 for teacher feedback and 1.13 for peer discussion combined with teacher feedback<sup>14</sup>, and show that students receiving feedback between  $Q_1$  and  $Q_2$  have considerably higher learning gains than students in the control condition who receive no feedback. Most improvement of learning gains in the short term occurs when peer discussion is

---

<sup>14</sup> The effect size here is defined as the difference between the mean of a treatment condition and the mean of the control condition on learning gains, divided by the pooled standard deviations for the two conditions (Lipsey et al., 2012). This method can be used since we randomly assigned students to classes and classes to treated and untreated conditions prior to the intervention. As a result of this, we expect to obtain an estimate of the average treatment effect on learning gains. We attempt to account for baseline differences by including covariates (such as basic demographic variables age and gender) that are both reliable and representative of learning gains (Rausch, Maxwell & Kelley, 2003).

immediately followed by teacher feedback. These results answer Porter et al.'s (2011) open question whether learning gains occur between  $Q_1$  and  $Q_2$  if, in addition to peer discussions, students also receive teacher feedback. Our results are similar to what is found in previous studies in science classrooms (Barth-Cohen et al., 2016; Smith et al., 2009; Smith et al., 2011; Zingaro & Porter, 2014). Additional analyses of our results show that student characteristics, time dummies, teacher characteristics and the level of difficulty of question pairs do not play a role in explaining the results of learning gain in the short term, on top of the differences due to feedback conditions.

The second research question explored whether learning gains are modified by peer discussions. The findings show that students in the cooperative condition who could learn from peer discussions (i.e. the students in the so-called potential learner group who answer  $Q_1$  incorrectly and  $Q_{1PD}$  correctly) are also likely to learn from peer discussions. These potential learners answer proportionally more  $Q_2$  correctly than students who understood the concept of the questions at the beginning. These findings are in line with the findings of Smith et al. (2009) and Egelandstad and Krumsvik (2017). We hypothesize that peer discussions help students to select the correct answer by prompting them to verbalize their solutions. Students learn from their peers and transfer acquired knowledge to new follow-up isomorphic questions. These findings are also in line with informal observations and conversations with students during our study. Students reveal that peer discussions break the monotony of passive listening to the teacher and encourage them to explain their own reasoning and listen to what others have to say. Peer discussions give them a sense of active participation and stimulate them to give a joint  $Q_{1PD}$  answer. Students appreciate it when peer discussions are immediately followed by teacher feedback. They emphasize the importance of this feedback in order to fully understand solutions or explanations when students are not sure of the correctness of their votes.

The findings in this study are generalizable to other educational settings. We firmly believe that the results can be applied to other secondary schools in the Netherlands, as our schools are representative for the average secondary school in the Netherlands. We showed that all statistics for these schools are within half of a standard deviation of the average of all variables. We also argue that the results are not content specific, as similar results have been demonstrated across several disciplines within secondary education and higher education,

such as psychology (Egelandsdal & Krumsvik, 2017b), biology (Knight et al., 2013; Smith et al., 2009; Smith et al., 2011), computer science (Porter et al., 2011; Zingaro & Porter, 2014), engineering statistics (Kjolsing & Van Den Einde, 2016) and physics (Barth-Cohen et al., 2016; Pollock, Chasteen, Dubson & Perkins, 2010). Furthermore, we used only question pairs that are isomorphic. To this purpose, pairs were randomized prior to the intervention and assessed for equality by two independent physics teachers, which implies that there is no reason to assume that there are learning gains from the fact that  $Q_2$  questions would be easier than  $Q_1$  questions. We also used the web-based program *Socrative*, which is very similar to the polling software *Kahoot!* and *TurningPoint*. For this reason, we assume that these polling programs will lead to the same results. Evidence using other software should confirm the generalization of our results.

As a general conclusion, it can be stated that teacher feedback, whether or not combined with peer discussions, improves learning outcomes. Although results suggest that peer discussion combined with teacher feedback increase learning outcomes in the short term, this study does not show to what extent these gains prompt students to retrieve information for the long term. Further studies are needed to evaluate the long-term effectiveness of peer discussions and teacher feedback when formatively assessed with SRS.



## Appendix 3.1

### Multilevel analyses – Robustness checks

As shown in Section 3.3.5, we do not use traditional regression analyses in which the student is the unit of observation. In our multivariate analyses, we aggregate all student responses (for each of the two treatment and control conditions separately), such that we can compare the conditions at the level of the question pair, and thereby account for the fact that not all question pairs are of the same level of difficulty. See Equation (3.1) in Section 3.3.5.

The database of 11,601 responses to pairs of questions has a hierarchical structure, as students are nested in classes and classes are nested in schools. It is therefore not inconceivable to use multilevel analyses. However, the consequences of using multilevel analyses is that we cannot account for the denominator in Equation (3.1). As such, we cannot control for the difficulty of the question pairs in a multilevel analysis and the variable  $LG$  is then equal (and limited) to  $(Q_2 - Q_1)$ . To check the robustness of our results of learning gains  $LG$  in Table 3.4, we apply multilevel analyses with our original dataset of 11,601 student responses (being aware that multilevel analyses rely on the incorrect assumption by not accounting for the difficulty of a question pair). The results of this robustness check with the outcomes learning gains  $LG$  with multilevel analyses are shown in Table 3.6. In Model A, we show a basic analysis that only includes the treatment status of the students, while in Model B and Model C we add covariates stepwise. In Model D, we estimate fixed effects by including time dummies to the analyses. The structure of Table 3.6 is identical to the structure of Table 3.4 in Section 3.4.1. Similar to the multivariate regressions in Models 1 through 4A in Table 3.4, the treatment conditions in the multilevel regressions in Models A through D in Table 3.6 are significant predictors of learning gains ( $\hat{\beta}_1 = 0.10$ ,  $p < 0.01$  and  $\hat{\beta}_2 = 0.18$ ,  $p < 0.01$ , respectively).

In Table 3.6, we also calculated the intraclass correlation coefficients (ICCs) to estimate the percentage of variance of the outcomes learning gains explained by unobserved student effects. In all models, we demonstrate that the percentage share of variance of learning gains explained by unobserved student effects is less than 0.05 ( $\rho = 0.021$  and  $\rho = 0.022$ ), which means that, according to Hox's (1998) rule of thumb, the use of multilevel analyses is not required.

By using multivariate analyses (Table 3.4) instead of multilevel analyses (Table 3.6), we consider the difficulty of the question pairs, while indirectly accounting for the different levels of observation. This justifies the fact that we use multivariate regressions throughout Chapter 3, in which we can control for the level of difficulty of the question pairs (and what is not possible in a multilevel regression).

Table 3.6: Multilevel regression analyses predicting learning gains

	Model A <i>Learning gains</i> <sup>A</sup>	Model B <i>Learning gains</i> <sup>A</sup>	Model C <i>Learning gains</i> <sup>A</sup>	Model D <i>Learning gains</i> <sup>A</sup>
( $\beta_1$ ) Individual condition	0.10*** (0.034)	0.10*** (0.034)	0.10*** (0.033)	0.10*** (0.034)
( $\beta_2$ ) Cooperative condition	0.18*** (0.031)	0.18*** (0.031)	0.18*** (0.031)	0.18*** (0.031)
( $\gamma_1$ ) Gender (boy = 0 & girl = 1)	--	0.0013 (0.0088)	- 0.0019 (0.0088)	- 0.0024 (0.0088)
( $\gamma_2$ ) Age	--	- 0.0011 (0.0060)	- 0.0055 (0.0061)	- 0.0055 (0.0061)
( $\gamma_3$ ) Education level	--	no	0.018 (0.025)	0.022 (0.026)
( $\gamma_4$ ) Physics test score prev. year	no	no	- 0.013*** (0.0047)	- 0.013*** (0.0047)
Fixed effects	no	no	no	yes time
Constant	0.15*** (0.026)	0.15*** (0.027)	0.14*** (0.031)	0.076** (0.035)
Observations	11,601	11,601	11,601	11,601
$\rho$	0.021	0.022	0.021	0.022

\*\*\* = significant at the 1 % level; \*\* = significant at the 5 % level; \* = significant at the 10 % level.

<sup>A</sup>The control condition is the reference category of the treatment conditions.



## Chapter 4

### The Effect of Feedback on Metacognition <sup>15</sup>

---

<sup>15</sup> This study is published in Computers & Education.

Reference: Molin, F., Haelermans, C., Cabus, S., & Groot, W. (2020). The effect of feedback on metacognition - a randomized experiment using polling technology. *Computers & Education*, 152, 103885.

This chapter explores the effects of formative feedback on students' metacognitive skills when using feedback strategies with student response systems (SRS). Using a randomized field experiment among 633 physics students in six schools in Dutch secondary education, we study assessments with the SRS *Socratic*, by dividing students into three conditions. Students in the cooperative condition use a combination of peer discussions and teacher feedback, while students in the individual condition use teacher feedback. To compare differences in metacognitive skills and motivation, students in the control condition only use *Socratic*, but do not receive formative feedback from either teacher or peers. The results show that there is a significant positive effect of the cooperative treatment on both metacognitive skills and motivation in comparison with the control condition. We find that students with low metacognitive skills benefit significantly more from the cooperative treatment than students with high metacognitive skills. No effects are found for the individual treatment. However, girls significantly increase their metacognitive skills and are more motivated than boys, when using an individual treatment. Additionally, a mediation analysis shows that motivation partially mediates the cooperative treatment and metacognitive skills. Based on these results, we recommend a combination of peer discussions and teacher feedback in physics courses.

## 4.1 Introduction

In daily teaching practice, formative feedback is a critical component of meaningful learning and the core of self-regulated learning, where students in dialogue with their teacher and peers are encouraged to monitor and regulate their own learning (Black & Wiliam, 1998a; Hattie & Timperley, 2007; Shute, 2008). Formative feedback helps students to clarify misunderstandings and identifies gaps in knowledge and skills. However, despite its importance, students in traditional classroom settings rarely receive the desired formative feedback, as teachers do not have enough time to provide formative feedback due to overloaded programs or congested classrooms, and sometimes simply lack the skills to assess students' understanding without grading (Lee et al., 2015; Trees & Jackson, 2007).

As stated in Chapter 1, researchers know that metacognitive skills and motivation are strongly and positively related to learning outcomes, and that formative feedback on students' understanding is indispensable in the learning process. For these reasons, it is crucial that teachers are aware of students' learning status and have the opportunity to provide formative feedback in a limited time and in a structured way. Previous literature shows that answering multiple-choice questions with SRS (e.g. clickers, *Socrative*, *Kahoot!*) creates opportunities for providing immediate formative feedback that meet these demands (Hunsu et al., 2016; Ludvigsen et al., 2015). By answering multiple-choice questions about the content that is being taught, students receive formative feedback and can monitor if the studied information is sufficiently understood. It enriches learning experiences by stimulating students to talk in classrooms about course content and it provides teachers with insight into students' thinking and learning (Tanner, 2009). Brady et al. (2013), Mayer et al. (2009) and Ludvigsen et al. (2015) pointed out that this kind of questioning in combination with formative feedback is a form of metacognitive monitoring; it gives students the opportunity to assess their knowledge and control or regulate their learning and performance. As multiple-choice questions are presented with content within contexts that have meaning to students, students acquire a broader repertoire of metacognitive monitoring activities, become more motivated and are more likely to use metacognitive skills. This is in line with studies that showed that students who receive immediate formative feedback have a higher intrinsic motivation to complete

their tasks (Kaddoura, 2013; Lin & Huang, 2018; Ryan & Deci, 2000) and a desire to acquire new metacognitive activities (DePasque & Tricomi, 2015; Edens, 2008).

Over the past decades, most SRS research has focused on the relationship between formative assessments and academic performance (Chien et al., 2016). Even though this research provides important insights, there is still a lack of evidence on why performance improvements among students occur. Based on findings in previous studies, we conclude that there is a need for more evidence how feedback affects students' metacognitive skills during formative assessments with SRS (Brady et al., 2013; Jones et al., 2012). Previous literature has indicated, for example, that students judge their state of learning during formative assessments based on the number of available cues that may be predictive of subsequent performance (Pyc, Rawson & Aschenbrenner, 2014; Tauber, Witherby & Dunlosky, 2019). To help students in identifying predictive cues, carefully designed interventions, such as metacognitive prompts are a requirement. These prompts enhance the utilization of diagnostic cues that are predictive of subsequent learning and performance (De Bruin, Dunlosky & Cavalcanti, 2017). The most frequently used metacognitive prompt during formative assessments with SRS is providing formative feedback, such as teacher feedback and peer discussions.

Only a limited number of quasi-experimental studies investigated students' metacognitive development in classes where SRS are used. Jones et al. (2012) reported that students' metacognitive awareness increases when students are formatively assessed with SRS. The use of SRS gives students multiple opportunities to receive immediate formative feedback and to try out their comprehension of the course material. The frequent formative feedback serves as a catalyst and is responsible for the improvement of students' regulation of cognition. Brady et al. (2013) argued that SRS use has a positive influence on the learning process and showed an increase of performance outcomes and metacognitive skills when SRS use is combined with instructional strategies and formative feedback. However, the studies of Brady et al. (2013) and Jones et al. (2012) did not use a proper randomized research design and did not have a control condition. Furthermore, these studies were done at university level with a limited number of students and groups.

In the present study, we use a randomized experiment to examine the effects of teacher feedback and peer discussions when using SRS on students' metacognitive skills and motivation during physics education in secondary schools. Each week, students receive conceptual multiple-choice questions facilitated with the online SRS *Socratic*. Students in the individual condition and cooperative condition answer the multiple-choice questions individually. To compare the influence of teacher feedback with a combination of peer discussions (cooperative condition) and only teacher feedback (individual condition), students in the cooperative condition discuss their responses in pairs (with peers) before the answer is presented and explained by the teacher. Students in the individual condition receive teacher feedback without having discussions in pairs. Students in the control condition only answer multiple-choice questions individually. As such, students in the control condition do not discuss responses in pairs, and receive no teacher feedback.

Our contribution to the literature is threefold: first, as far as we know, the study at hand is one of the first studies that examine the effects of teacher feedback, whether or not combined with peer discussions, using SRS on metacognitive awareness and motivation. We do so by conducting a randomized controlled trial (RCT) with sufficient statistical power. Previous studies are of quasi-experimental nature with a pre- and post-test design, but without a control condition (e.g. Brady et al., 2013; Edens, 2008; Jones et al., 2012; Zhonggen, 2017). A second contribution is that we compare two treatment conditions, with different types of feedback, with a control condition, enabling us to compare differences in metacognitive awareness and motivation. Furthermore, with an interaction analysis we are able to show differential effects on heterogeneous populations, for example, we estimate the effects for girls and boys separately, and also across pre-treatment levels of metacognitive awareness of students. Lastly, as we measure both metacognitive skills and motivation, we are able to show the mediating effect of motivation on metacognitive awareness of students.

The effectiveness of each feedback strategy is evaluated for the outcome metacognition and the intermediate outcome motivation. Therefore, our research questions are:

1. What is the effect of teacher feedback or peer discussions combined with teacher feedback on metacognition?



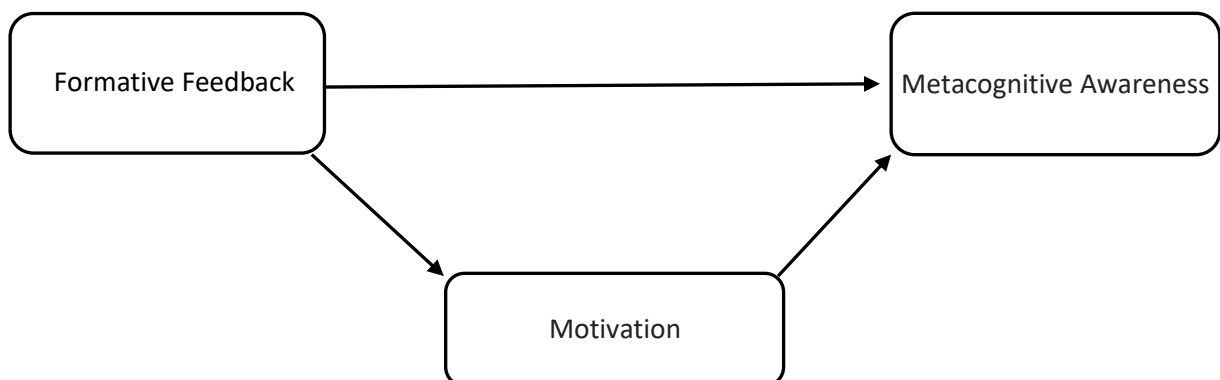
2. What is the effect of teacher feedback or peer discussions combined with teacher feedback on motivation?
3. Are there differential effects among subgroups of students?
  - a. Are the effects different for girls and boys?
  - b. Are the effects different for students with high metacognitive skills, students with middle metacognitive skills, or students with low metacognitive skills?
4. Does the effect of teacher feedback, or peer discussions combined with teacher feedback, on metacognition run (partly) through its influences on motivation?

This chapter continues with a conceptual framework and an overview of the literature in Section 4.2, followed by an overview of the experiment and the data and descriptive statistics in Section 4.3. In Section 4.4 we discuss the results. Section 4.5 concludes the chapter and discusses the findings.

## 4.2 Conceptual framework and literature

Based on the literature pertaining to metacognition, motivation and formative feedback during SRS use, we have developed a conceptual framework that we will test in the empirical part of this chapter. The conceptual framework identifies the key factor formative feedback. On the one hand, formative feedback may directly affect metacognitive awareness. On the other hand, formative feedback can influence metacognitive awareness and skills indirectly through its influences on motivation (Figure 4.1).

Figure 4.1: Model of relations



Metacognition is defined as one's ability to regulate cognitive processes and is strongly related to learning outcomes (Coutinho, 2006; Flavell, 1979; González et al., 2017; Thomas, 2013). It refers to higher-order thinking, and includes skills that enable learners to think about, understand, and monitor their learning (Schraw & Dennison, 1994). Students with more metacognitive awareness and skills are better equipped to take advantage in learning environments. Their learning allows them to monitor and regulate cognitive activities in a way that improves learning outcomes (Nelson & Narens, 1990). A central aspect here is self-regulated learning, which is a process whereby students set goals for their learning and where they monitor and control actions, cognition and motivation needed to achieve these goals (Zimmerman & Labuhn, 2012). A general belief is that motivation is an important factor in monitoring and regulating learning processes, because motivation stimulates students' learning by focusing attention on the task, which results in an improvement of metacognitive awareness and learning outcomes (Pekrun, 2006; Pintrich & de Groot, 1990). Focusing attention on a task is a consequence of an intrinsic motivation to perform that task (Eccles, 2005). A lack of focusing attention implies that intrinsic motivation, or the motivation to work on a task primarily for its own benefit, is lacking and that extrinsic motivation, or the motivation to engage in a task because it is a means to an end, is necessary to sustain the focus of attention on the task (Deci et al., 1996; Pekrun et al., 2010; Sansone & Thoman, 2005). This might happen when the feedback received implies a lack of competence of the student, which reduces the interest and subsequent choice to perform the task. Intrinsic and extrinsic motivation are intertwined when performing tasks (Sansone & Thoman, 2005). In this chapter we capture the terms *intrinsic motivation* and *extrinsic motivation* with the umbrella term 'motivation', see also Figure 4.1.

In daily teaching practice, formative feedback is the core of self-regulated learning, where students in dialogue with their teacher and peers are encouraged to monitor and regulate their own learning. This helps students to broaden and deepen their own learning process outcomes (Butler & Winne, 1995; Shute, 2008; Winne & Hadwin, 1998), and changes students' attention and motivation during learning, which contributes to the satisfaction of students' basic needs to feel competent (Chien et al., 2016; Ryan & Deci, 2000). Mazur (1997) and Barth-Cohen et al. (2016) asserted that formative feedback operationalized through a questioning-integrated instruction (e.g. quizzing) stimulates students to monitor their learning and control

or regulate their learning process. Quizzing can be implemented in classrooms in a high-tech setting, where students are allowed to enter their answers into some kind of device, for example clicker devices (Blasco-Arcas et al., 2013; Mayer et al., 2009), the web-based system *Socrative* (Balta & Tzafilkou, 2019; Kokina & Juras, 2017) or *Kahoot!* (Wang, 2015). These SRS rapidly collects, records, and displays the number of correct and incorrect answers on a screen in front of the class. Compared with feedback in low-tech settings, where students indicate their answer option by simply raising their hands, raising coloured flashcards, applause or showing mini whiteboards, the teachers' feedback in a high-tech setting can be more authentic and is less threatening to students' self-esteem, because their performance is evaluated in a more private way (Caldwell, 2007; Kay & LeSage, 2009).

Assessments with SRS offer students the opportunities to receive immediate formative feedback in a single lesson, several times, and in many different ways. The most basic way of providing feedback is simply to show a checkmark next to the correct answer while the teacher reads aloud the correct answer. The only feedback students receive here is whether or not they correctly answer the question, no more and no less. The study of Lantz and Stawiski (2014) showed that assessments with this limited amount of feedback slightly positively increases final test scores compared to conditions receiving no multiple-choice questions and no feedback throughout the lecture. In a more extensive way of SRS use, the high-tech system aggregates all answers of the entire class and provides formative feedback by showing voting statistics of students' responses (e.g. in the form of a bar chart, pie chart or histogram) in front of the class with a checkmark next to the correct answer. Here, the voting statistics are a source of information for students to monitor their own knowledge with reference to their fellow students. It allows students to realize that they are not alone in struggling with course material if they gave an incorrect answer, which reduces self-doubt and the feeling of whether they are incapable of understanding the content (Chien et al., 2016; Hoekstra & Mollborn, 2012; Kay & LeSage, 2009; Knight & Wood, 2005). The graphical representation of students' responses also makes feedback accessible for teachers, giving them the opportunity to explain the correct answer and inform students about which common mistakes are made. Bachman and Bachman (2011), Bartsch and Murphy (2011) and Mayer et al. (2009) all found that showing a distribution of answers and receiving teacher feedback, on top of only giving the correct answer, has a positive impact on student learning and leads to higher scores on tests

and exams. However, the studies of Ludvigsen et al. (2015), Evans (2013) and Vickrey et al. (2015) pointed out that teacher feedback alone might be insufficient for students. It is possible that students find interpretations of subject-specific terms problematic and thereby they experience difficulties when they have to apply these in their own learning processes. In a more extended way of providing formative feedback, teacher feedback can be more comprehensible for students when it is combined with peer discussions. Peer discussions are formative in nature and belongs to a cooperative learning strategy that assesses one's knowledge in a more accessible language than teacher feedback alone (Birenbaum, 1996; Gielen, Tops, Dochy, Onghena & Smeets, 2010). The reason for this is that students have a similar background and use a similar language; they can explain problems and solutions in different ways than a teacher, without using subject-specific terms (Blasco-Arcas et al., 2013; Caldwell, 2007; Perez et al., 2010). Peer discussions also encourage students to explain and justify their own reasoning or interact with peers to arrive at an answer (Chien et al., 2016; Cortright, Collins & DiCarlo, 2005; Levesque, 2011; Smith et al., 2009). In a common strategy of peer discussions, students first answer a multiple-choice question individually, then discuss their reasoning with their peers, and finally answer again (and potentially change their answer, based on the discussion with their peers) before the outcome to the question is shown and explained (Crouch & Mazur, 2001; Smith et al., 2009; Vickrey et al., 2015). This form of explanation stimulates students to retrieve, combine and adjust their own existing knowledge with new knowledge. Blasco-Arcas et al. (2013), Brady et al. (2013), Levesque (2011) and Mazur (1997) reported that multiple-choice questions supported by peer discussions affect metacognitive monitoring and develop metacognitive skills for determining how well students understand course material and how to solve problem-like questions in the future.

Formative feedback (for example provided by teachers or peers) regarding aspects of one's understanding can trigger motivation (Figure 4.1). Especially when this feedback helps students to compare learning goals to their own activities; it motivates them to reduce the discrepancy between what is understood and what is aimed to be understood (Hattie & Timperley, 2007). The studies of Camacho-Miñano and Del Campo (2016) and Sun and Hsieh (2018) showed that providing formative feedback and active learning through assessments with SRS scaffold the development of students' motivation. The SRS stimulate the teacher-student or student-student interaction, enabling the teacher to make lessons more interactive

and interesting and students more proactive and focused when answering. The anonymity of SRS, where chosen answers of individual students are not publicly revealed to all students in the classroom and not immediately apparent to teachers, is also appealing to students. It reduces students' anxiety (Yu et al., 2014) and provides more involvement and active participation of students in the classrooms (Stowell & Nelson, 2007; Zhonggen, 2017). This anonymity may in particular stimulate insecure students to actively participate as well, by digitally answering multiple-choice questions and contributing to class discussions, giving them more time to process information (Bartsch & Murphy, 2011; Bojinova & Oigara, 2013). In short, interactive activities, like assessments with SRS, stimulate students to do more than passively listen to the teacher, provide immediate feedback, encourage monitoring by giving students the opportunity to assess their knowledge, and increase students' motivation, which finally all leads to more metacognitive awareness (Buil et al., 2016; Caldwell, 2007; Edens, 2008; Tlhoale, Hofman, Winnips & Beetsma, 2014).

Previous literature has shown that quizzing with SRS with immediate formative feedback has the potential to increase students' metacognitive awareness and motivation. However, more research is needed on (a) to what extent an individual or a cooperative treatment affects metacognitive awareness and motivation; and (b) whether these feedback strategies benefit some students more than others. Some researchers suggest that there is a connection between teacher feedback or peer discussions and growth of metacognitive awareness/skills and motivation on the one hand, and teacher feedback or peer discussions and gender on the other hand (e.g. Brady et al., 2013; Jones et al., 2012; Kang, Lundeberg, Wolter, delMas & Herreid, 2012; Mayer et al., 2009). One of the few interventions that reports about SRS use and the growth of self-knowledge is Shapiro et al. (2017), who found that students who are lacking deep motivation and have low metacognitive skills, benefit most from peer discussions during assessments with SRS. Shapiro et al. (2017) stated that multiple-choice questions boost low-metacognitive aware students to a higher level equal to their more motivated and high self-knowledge peers. The researchers also found a reduced magnitude effect for students with high metacognitive skills, suggesting that the used multiple-choice questions required less problem solving skills.

As to feedback and gender, King and Joshi (2008) suggested that activities with SRS use does not discriminate gender, but it might be easier for one gender to participate. Previous

literature stated that both boys and girls benefit from interactive learning strategies (Lorenzo, Crouch & Mazur, 2006; King & Joshi, 2008), but that they approach learning differently (Brotman & More, 2008). Boys are more competitive and less focused on the quality of feedback, while girls prefer learning when they can express their ideas through discussions (Lorenzo et al., 2006; King & Joshi, 2008).

## 4.3 Materials and methods

### 4.3.1 The intervention

To test our conceptual framework of Figure 4.1 and answer our research questions, we set up an intervention that consists of assessments with SRS in two treated conditions (an individual and a cooperative condition) and one untreated condition (a control condition). Students in the cooperative condition use a combination of peer discussion and teacher feedback, while students in the individual condition only use teacher feedback. The students in all three conditions answer on average eight multiple-choice questions per week for a period of 10 weeks, divided into four sets of two paired questions. These paired questions assess the same conceptual understanding and the same concepts or principles, but in different contexts and with different numerical values (Smith et al., 2009). See Figure 3.2 in the previous chapter for two examples of two paired questions. All weekly questions cover a part of the course content and are selected for higher-order thinking skills on *applying* and *analyzing* of Bloom's taxonomy (Bloom et al., 1956), by two physics teachers who have experience in using this taxonomy.<sup>16</sup> Using questions on *applying* and *analyzing* enhances student study skills and metacognitive awareness as it helps them monitor their mastery of the learning material (Bloom et al., 1956; Crowe, Dirks & Wenderoth, 2008). Each multiple-choice question has four possible answers, of which only one is correct. Because almost all students have their own smartphone, we use the online SRS *Socrative*, which runs on all mobile digital devices (like smartphones, laptops and tablets) that can access the internet. Students who did not have a smartphone, or whose battery was empty, could borrow a laptop from school. This only

---

<sup>16</sup> There are six levels in Bloom's taxonomy, each requiring a higher level of abstraction from students: (1) *knowledge level* (remembering of previously learned material), (2) *comprehension level* (understanding the meaning of material), (3) *application level* (ability to use material in new situations), (4) *analysis level* (ability to see patterns that can be used to analyze problems), (5) *synthesis level* (ability to put parts together and make new predictions or create new theories), and (6) *evaluation level* (ability to judge the value or bias of material).

occurred a few times. Therefore, it should not have interfered with the general effects of using SRS on smartphones as the main device.

The questions are projected both on a large screen in front of the class and on students' own smartphones (see Figure 4.7 in Appendix 4.1 of two screenshots of a paired question as students see them on their smartphones). Depending on the nature and difficulty of the questions, students are given 2 - 4 minutes to answer a question individually using their smartphones, with no interactions between them. After voting, the answers are collected by the *Socrative* software. Here, all question responses of all students, even null responses, are recorded by the *Socrative* software. Because students' names are not attached to responses, students stay anonymous for the teacher. The responses are not graded and do not affect students' final grades, so students can vote without concern for whether they are correct. Although this method of *Socrative* use is the same for all three conditions, the manner in which feedback is provided is different, as is described below.

#### *4.3.1.1 Control condition*

For students in the control condition, the feedback they receive is only seeing the correct answer of a multiple-choice question in the form of a checkmark. They receive no feedback on voting statistics of students' responses and get no explanations of correct and incorrect answers from their teacher or peers (Figure 4.2). We address the ethical dilemma, of having a control condition that does not receive any benefit, by giving all participating students in this study the opportunity to access a protected website to voluntarily read the limited overall feedback of all correct and incorrect answers, after class. However, we have no information about whether students made use of this on a regular basis.

#### *4.3.1.2 Individual condition*

Students in the individual condition answer the same multiple-choice questions in the same time span as the control condition with their smartphones individually, introducing the difficulty level of the questions across the treated and untreated conditions in a similar way. While students in the control condition only see a correct answer, for students in the individual condition the responses of a multiple-choice question are displayed anonymously as a histogram with a distribution of answers on a screen in front of the class. The checkmark

informs the students of the correct answer, and the teacher explains the answer and informs them on which reasoning errors underlie incorrect answers (Figure 4.2).

#### *4.3.1.3 Cooperative condition*

After voting a multiple-choice question, in the cooperative condition the teacher does not immediately reveal the correct answer, and the histogram of students' votes is also not shown. Here, the multiple-choice question is set to appear twice and supported by a peer instructional approach (Mazur, 1997). Students are given a limited time of 2 - 3 minutes to discuss their individual responses and alternative answers with their peers (in pairs), after which they answer the same question for the second time individually. In the meantime, the teacher walks around in the classroom, listens to the students' peer reasoning and interacts with them, stimulating critical discussions (focused on the reasons behind their answers instead of only discussing answers). Then students answer the same question again, immediately followed by a histogram of the latest responses. A checkmark confirms the correct answer, after which the teacher explains the correct and incorrect answers (Figure 4.2).

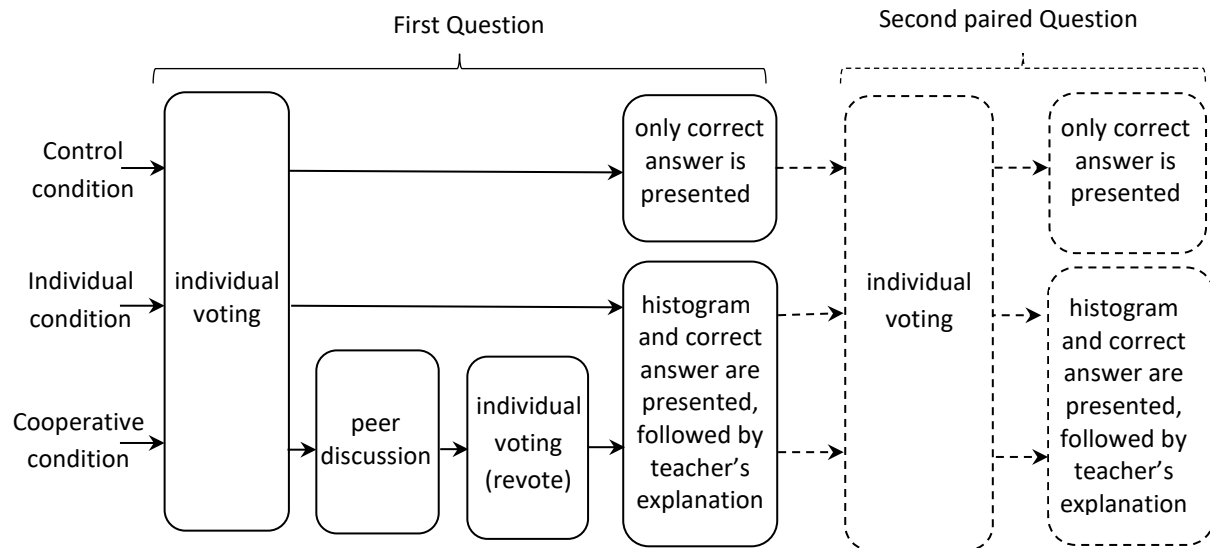
As in the control and individual conditions, students are also asked to respond to a second follow-up question individually (dashed blocks in Figure 4.2). After voting this question, a checkmark informs the students of the correct answer. The teacher in a cooperative and individual condition shows the histogram of student responses and gives students immediate formative feedback on how well the concepts are understood. The process of asking paired questions shown in Figure 4.2 is repeated approximately four times per week in each condition.

The design of this study is similar to the method of Jones et al. (2012), where students with a cooperative feedback strategy (peer discussions combined with teacher feedback) and students with an individual feedback strategy (only teacher feedback) answer multiple-choice questions with SRS weekly. However, in contrast to Jones et al. (2012), we also use a control condition and carry out an intervention with two treatment conditions that can be compared with this control condition. Additionally, we use pairs of questions. Answering these questions stimulates students to convince their peers of the correct answer and goes beyond just merely copying correct answers from teachers or peers. It gives students the opportunity to reflect



upon their problem-solving processes and develop reasoning skills and metacognitive skills (Smith et al., 2011; Zingaro & Porter, 2014). As shown in Chapter 3, students could learn from teacher feedback and peer discussions and are better able to answer correctly a second conceptual-related question.

Figure 4.2: Experimental design



For this experiment, we have created a database with 450 paired sets of multiple-choice questions. Each pair of questions is designed to exercise the same level of conceptual understanding and has the same level according to Bloom's Taxonomy; namely the application-level or analysis-level (Bloom et al., 1956). This database then consists of 265 paired sets of questions that belong to the application-level, and 185 paired sets of questions that belong to the analysis-level. The sets of multiple-choice questions cover content of the Dutch upper secondary school physics syllabus (*mechanics, waves, thermodynamics, electromagnetism and modern physics*). Depending on the topic of the week, a teacher selects four paired sets of multiple-choice questions (eight questions in total) that students answer using *Socrative*.

### 4.3.2 Participating schools and students

In this study, over six hundred students from six secondary schools in the southern part of the Netherlands participate. All schools are located in the provinces of Limburg and Brabant, outside the highly urbanized region of the Netherlands and are representative schools for Dutch secondary education based on student numbers, employed teachers and performance.

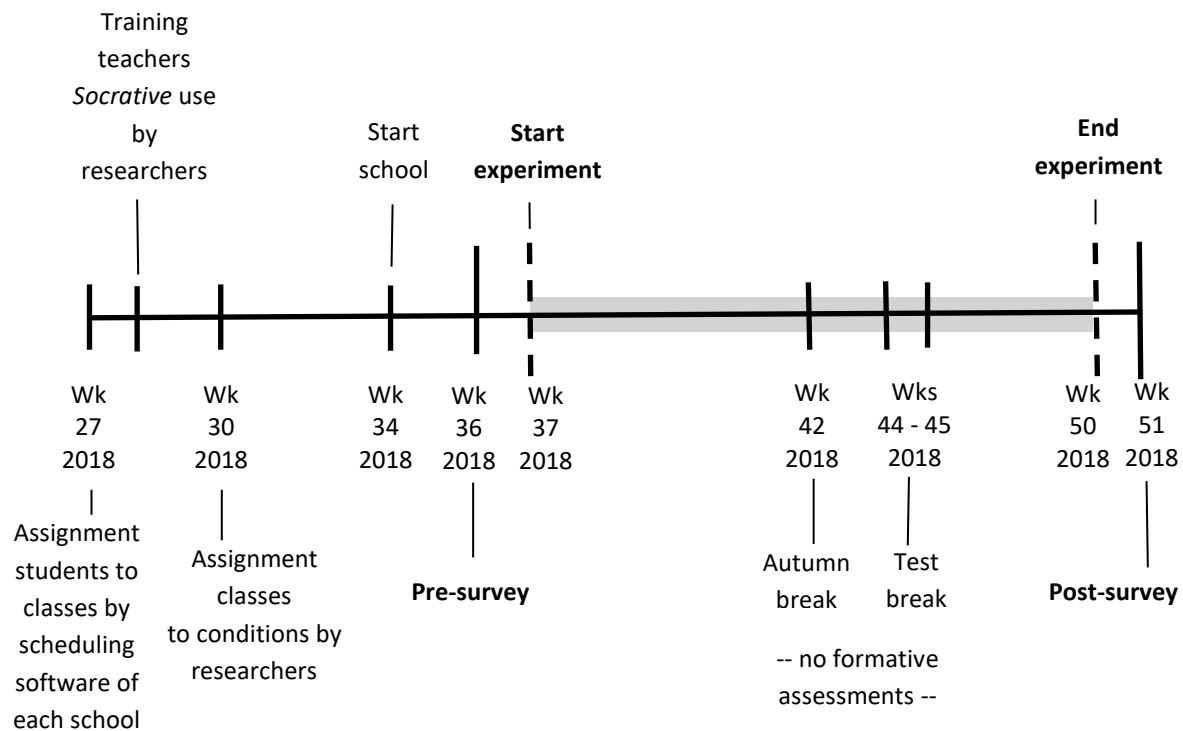
Thirteen physics teachers participated in this study. Seven hundred and forty-one physics students in thirty-three classes, in grade 10 - 12 are part of the experiment. Despite the fact that all 741 students from these teachers' classes completed the pre-survey at the start of the experiment, 108 students are excluded from the analysis due to not completing the post-survey at the end of the experiment ( $N = 32$ ), leaving school or class during the experiment ( $N = 7$ ) or hardly participating in the weekly assessments ( $N = 69$ ) and therefore not sufficiently meeting the intervention conditions. As a result, the final sample of this study comprises 633 students nested within nine classes with the individual condition, fifteen classes with the cooperative condition and nine classes with the control condition. To check whether the students are a specific group with respect to one or more of the background characteristics, we also carry out  $t$ -tests between included ( $N = 633$ ) and excluded ( $N = 108$ ) students. Here, no significant differences are shown in student characteristics between groups. See Table 4.6 in Appendix 4.2 for the  $t$ -tests;  $p > 0.05$  in all tests.

### 4.3.3 Timeline of the experiment

A few months before the start of the experiment, six schools agreed with the treatment conditions of the treated and untreated groups, after which the physics teachers involved were informed about the aims of this study. To increase proficiency, and to ensure that assessments and the use of the technology would be used correctly, the teachers received a short training session by the researchers, on the use of *Socratic* and the preconditions under which the treatments have to occur (Figure 4.3). They also received the database with multiple-choice questions for creating the *Socratic* quizzes. In the meantime, the students were randomly assigned to classes by the scheduling software and three weeks later classes were randomly assigned to either one of the treatment conditions or the control condition by the researchers. The experiment took place over a period of thirteen weeks, but lasted only ten weeks in total, due to a one-week autumn holiday and a two-week testing period in

between (the exams in this testing period were not part of the intervention). One week before the experiment, a pre-survey on motivation and metacognitive skills was taken, and one week after the experiment, a post-survey was taken. Note that the literature indicates that a time span of ten weeks should be long enough to measure a change in metacognitive skills (Andersson & Palm, 2017; Jones et al., 2012).

Figure 4.3: Overview of the timeline of the experiment



#### 4.3.4 Identification strategy

We use a randomized experimental design including six participating schools, which allows for causal analysis and increases the internal validity of this study. To do this, the students of each school are randomly assigned by the scheduling software to one of the physics classes, conditional on their education level. The randomizations take place before the timetables are made. Next, the classes are randomly assigned within schools by throwing a dice to the treated and untreated conditions, again conditional on education level. By randomizing classes in this way, we may assume that we minimize possible teacher effects, as selection into treatment and control condition was random and did not depend on the teacher. By including six schools in the experiment, and by randomizing within schools, we minimize possible school effects and increase the external validity of our study.

Furthermore, we minimize the risk of contamination, i.e. when students share information about discussions of multiple-choice questions outside the classroom. To avoid this potential source of bias, students are obliged to make their notes on school paper and leave these in the classroom at the end of each lesson.

#### 4.3.5 Measurement instruments

Previous literature shows that it is not easy for students to express their thoughts about their own metacognitive awareness and motivation, as these are internal processes that students are often unaware of (Avargil, Lavi & Dori, 2018; Desoete, 2008; McCombs, 1996). It is often difficult for researchers to measure such psychological constructs (Brady et al., 2013; Panaoura & Philippou, 2005; Schraw & Impara, 2000). Therefore, we use valid and reliable questionnaires for assessing metacognitive awareness and motivation.

##### 4.3.5.1 Metacognition

Students' metacognition is assessed by using the *Metacognitive Awareness Inventory* (MAI; Schraw & Dennison, 1994). This validated, self-reported questionnaire assesses metacognitive skills of students. It contains 52 items that examine '*what someone knows about learning and about oneself as a learner*' (knowledge of cognition), and '*the skills of monitoring and activities of someone that helps control one's thinking and learning*' (regulation of cognition). Students respond on a five-point Likert scale ranging from 'totally disagree' to 'totally agree'. The total score for the MAI is the average of the scores of all five-point Likert scale items, where 1 is the minimum score and 5 the maximum. A higher score indicates a higher level of metacognitive awareness and skills (Pintrich, 2000b; Schraw & Dennison, 1994).

##### 4.3.5.2 Motivation

Students' motivation towards learning physics is assessed by completing the motivation scale of the *Physics Motivation Questionnaire* (PMQ; Glynn & Koballa, 2006), consisting of 10 five-point Likert scale items ranging from 'never' to 'always'. The questionnaire conceptualizes students' motivation to learn physics in terms of the two sub-scales (intrinsic motivation and extrinsic motivation) and is translated into Dutch. The minimum average score of the ten items of PMQ is 1 and the maximum 5. Higher scores represent more motivation to learn physics (Glynn, Taasobshirazi & Brickman, 2009).

The items of the PMQ and MAI are both translated into Dutch by two independent qualified English teachers; both native speakers in English and proficient in Dutch. After translation, the two teachers met each other and compared their work with the original English versions. The teachers resolved inconsistencies after in-depth discussions and generated joint Dutch versions. In order to check the consistency with the original English questionnaires, the Dutch questionnaires were translated back into English and compared with the original versions. Finally, the Dutch questionnaires were distributed to 18 students, who were not included in the study. They completed the questionnaires and commented on the items in the event of ambiguities.

#### 4.3.5.3 Pre-survey information

Before the start of the experiment, all students were asked to complete a pre-survey which provides demographic data (gender, age, physics test score of previous school year and *Socratic* use in previous school year) and the pre-tests *Metacognitive Awareness Inventory* (MAI; Schraw & Dennison, 1994) and *Physics Motivation Questionnaire* (PMQ; Glynn & Koballa, 2006). The pre-survey was conducted in each school and in each class in week 36; one week before the experiment started (Figure 4.3). The teachers explained the purpose of the survey and asked students to take part in this, having said that this was voluntary. All present students voluntarily completed the survey in thirty minutes in the presence of the teacher. Finally, all students that were absent during that class voluntarily conducted the pre-survey the same week during class.

For the pre-MAI, the degree of internal consistency, measured with Cronbach's alpha for all participating students is good (Cronbach, 1951):  $\alpha = 0.90$ . This alpha is higher than the 'critical' value of 0.70 (Tavakol & Dennick, 2011). The same applies to the sub-scales 'knowledge of cognition' (17 items:  $\alpha = 0.78$ ) and 'regulation of cognition' (35 items:  $\alpha = 0.88$ ). All Cronbach alphas are similar to earlier reported values of the MAI (e.g. Jones et al., 2012; Schraw & Dennison, 1994).

The coefficient of internal consistency for the pre-PMQ of all participating students is acceptable with Cronbach's alpha equal to 0.74. The same applies to the sub-scales 'intrinsic motivation' (5 items:  $\alpha = 0.70$ ) and 'extrinsic motivation' (5 items:  $\alpha = 0.70$ ) which all

corresponds to the findings of Glynn and Koballa (2006) and Glynn et al. (2009). Based on the ‘critical’ value of 0.70 (Tavakol & Dennick, 2011), from the (aforementioned) Cronbach’s alphas we conclude that the pre-survey instruments are valid and reliable.

#### 4.3.5.4 Post-survey information

At the end of the experiment, students’ metacognition and motivation were assessed by completing the same survey, following the same procedures. This post-survey was conducted in the last week before the Christmas break (Figure 4.3).

The Cronbach alpha dealing with the responses on the post-MAI is again good ( $\alpha = 0.90$ ). We argue that the post-survey instruments stay valid and reliable and comparable to the pre-survey instruments (knowledge of cognition:  $\alpha = 0.79$ ; regulation of cognition:  $\alpha = 0.88$ ), as well as acceptable for the responses for the motivation scale of the post-PMQ ( $\alpha = 0.76$ ), intrinsic motivation ( $\alpha = 0.76$ ) and extrinsic motivation ( $\alpha = 0.70$ ).

#### 4.3.6 Descriptive statistics

Table 4.1 summarizes the descriptive statistics of students involved in this study. In total, 633 students from 33 different classes participate in this study, of which 47 % are girls. None of the students reported a non-binary gender. For this reason we compared boys versus girls. At the start of the intervention, the students are on average 15.7 years old, although age ranges from 14 to 19. The average physics test score of previous year is 6.55 (scored on a scale between 1.0 and 10.0). Seventy-eight percent of the students are familiar with the use of *Socratic*. The average pre-score of metacognition is 3.28 ( $SD = 0.37$ ), scored on a scale between 1 and 5. The average pre-score for motivation is 3.44 ( $SD = 0.54$ ), again scored on a scale between 1 and 5.

Table 4.2 presents a comparison of the metacognitive and motivational skills of the control condition, the individual condition and the cooperative condition. The quality of the randomization was examined using analysis of variance (ANOVA). The ANOVA tests show that students in the two treatment (individual- and cooperative) conditions are, on average, similar to students of the control condition.

In Section 4.4, we control for all of these student pre-treatment characteristics in our regressions. Here, we will only use standardized (z)-scores of all components in order to facilitate easy interpretation of the results.

Table 4.1: Descriptive statistics of the sample

	<i>N</i>	<i>Mean</i>	<i>Std. Dev.</i>	<i>Min.</i>	<i>Max.</i>
Age	633	15.73	1.03	14	19
Physics test score <sup>A</sup>	633	6.55	1.00	1.0	10.0
<i>Socratic</i> use <sup>A</sup>	633	0.78	0.42	0	1
Metacognition <sup>B</sup>	633	3.28	0.37	1	5
Motivation <sup>B</sup>	633	3.44	0.54	1	5

<sup>A</sup> previous year

<sup>B</sup> average pre-score

Table 4.2: ANOVA results for pre-tests between conditions

	<i>Control condition</i> ( <i>N</i> = 190)		<i>Individual condition</i> ( <i>N</i> = 179)		<i>Cooperative condition</i> ( <i>N</i> = 264)		<i>F</i>	<i>p</i>
	<i>Mean</i>	<i>Std. Dev.</i>	<i>Mean</i>	<i>Std. Dev.</i>	<i>Mean</i>	<i>Std. Dev.</i>		
Gender <sup>A</sup>	0.48	0.50	0.42	0.49	0.50	0.50	1.37	0.25
Age	15.65	1.10	15.75	0.93	15.77	1.04	29.46	0.44
Physics score <sup>B</sup>	6.50	1.09	6.53	0.77	6.59	1.07	0.52	0.59
<i>Socratic</i> use <sup>B</sup>	0.79	0.40	0.74	0.44	0.78	0.41	0.80	0.45
Metacognition <sup>C</sup>	3.26	0.40	3.27	0.39	3.30	0.35	1.01	0.37
Motivation <sup>C</sup>	3.44	0.56	3.45	0.51	3.45	0.54	0.15	0.86

\*\*\* = significant at the 1 % level; \*\* = significant at the 5 % level; \* = significant at the 10 % level.

<sup>A</sup> boy = 0, girl = 1

<sup>B</sup> previous year

<sup>C</sup> average pre-score

#### 4.3.7 Methods of analyses

In the present study, we use analysis of variance (ANOVA) to test whether an individual or cooperative feedback strategy affects metacognition and motivation. Here, the ANOVA determines whether the means of the independent variables of our treated and untreated conditions significantly differ from each other. To identify the outcomes, we observe a student  $i$  clustered within a class  $j$ , with the intervention being performed at the class level.

The regression can be formulated as follows:

$$Y_{ij} = \alpha_0 + \beta_1 T_{1,ij} + \beta_2 T_{2,ij} + \sum_{k=1}^k \gamma_k X_{k,ij} + \nu_j + \varepsilon_i \quad (4.1)$$

where  $Y_{ij}$  denotes the outcome variable metacognition or motivation of a student  $i \in \{1,2,\dots,633\}$  attending physics class  $j \in \{1,2,\dots,33\}$ ,  $T_i$  is the treatment status of a student  $i \in \{\text{individual } (T_1); \text{cooperative } (T_2)\}$ ,  $X_i$  is a vector of students' observable pre-treatment characteristics which are independent of the treatment conditions (such as metacognition ( $\gamma_1$ ), motivation ( $\gamma_2$ ), gender ( $\lambda_1$ ), physics test score of previous year ( $\gamma_3$ ), age ( $\gamma_4$ ) and *Socratic* use previous year ( $\gamma_5$ )) and  $\nu_j$  and  $\varepsilon_i$  are the error components at class level and student level, respectively. Because we randomized at the class level, all standard errors in the analyses of this study are clustered at the class level. By doing this, we correct for internal correlations between characteristics that may be common to students who attend lessons in the same classes (such as teacher, class composition, and class environment) (Moulton, 1986). As the number of clusters in this experiment ( $n = 33$ ) exceeds the minimum amount of around 30 that is usually expected to apply clustered standard errors at the class level without the risk of bias (Angrist & Pischke, 2008; Haelermans & Ghysels, 2017; Wooldridge, 2010), using clustered standard errors is appropriate and accepted. By including several observed student characteristics to the analyses, we increase the precision of our estimates.

To analyze our third research question, we also estimate two additional specifications by extending our previous regression. To determine whether the treatment has heterogeneous effects by gender, we include in Equation (4.2) an interaction between the treatment conditions and gender:

$$Y_{ij} = \alpha_0 + \beta_1 T_{1,ij} + \beta_2 T_{2,ij} + \lambda_1 \text{Gender}_{ij} + \theta_1 (T_{1,ij} \times \text{Gender}_{ij}) + \theta_2 (T_{2,ij} \times \text{Gender}_{ij}) + \sum_{k=1}^k \gamma_k X_{k,ij} + \nu_j + \varepsilon_i \quad (4.2)$$

To contrast the effect of the treatment condition for students with high metacognitive skills to other metacognition level groups, we include in Equation (4.3) an interaction between treatment condition and subgroups of metacognitive level. In this interaction analysis, the sample of students is divided into three roughly equal-sized subgroups (low-, middle- and



high metacognitive skills), based on the metacognition score of the pre-test (the overall MAI-score). The low-metacognitive skill group includes the 33.3 % lowest-scoring students on pre-metacognition, while the high-metacognitive skill group includes the 33.3 % highest-scoring students on pre-metacognition. The remaining 33.3 % students belong to the middle-metacognitive skill group. Note that the three subgroups of the individual, cooperative and control condition are not exactly equal in size. The reason for this is that we use complete scores as cut-off points<sup>17</sup>, after which we rearrange the sample of students to the three treatment conditions. We then may write:

$$\begin{aligned}
 Y_{ij} = & \alpha_0 + \beta_1 T_{1,ij} + \beta_2 T_{2,ij} + \lambda_2 \text{Low Metacog}_{ij} + \lambda_3 \text{Middle Metacog}_{ij} + \\
 & \eta_1 (T_{1,ij} \times \text{Low Metacog}_{ij}) + \eta_2 (T_{1,ij} \times \text{Middle Metacog}_{ij}) + \\
 & \eta_3 (T_{2,ij} \times \text{Low Metacog}_{ij}) + \eta_4 (T_{2,ij} \times \text{Middle Metacog}_{ij}) + \sum_{k=1}^k \gamma_k X_{k,ij} + V_j + \varepsilon_i
 \end{aligned} \tag{4.3}$$

Lastly, to test for mediation, we use a series of four regression analyses described by Baron and Kenny (1986). We analyze if (1) there is a significant effect between the treatment and the presumed mediator motivation, (2) the treatment and (3) the presumed mediator motivation are significantly related to the outcome variable metacognition, and (4) the effect of teacher feedback or peer discussions combined with teacher feedback on metacognition is reduced (partial mediation) or no longer significant (complete mediation) if we control for the presumed mediator motivation.

## 4.4 Results

### 4.4.1 Results metacognition

Table 4.3 summarizes the effects of teacher feedback (individual condition) or peer discussions combined with teacher feedback (cooperative condition) on metacognition. All models are estimated using regression analysis with clustered standard errors. The first model is a basic model with outcome metacognition that only includes the treatment status of the students. As compared to the control condition, Model 1 indicates a positive effect of

---

<sup>17</sup> We could also opt for different methods to classify the three subgroups, for example by using cut-off percentiles at 15.9 and 84.1 corresponding to the first standard deviation of normal distribution. However, if we choose other cut-off points our results in Section 4.4.1 do not change.

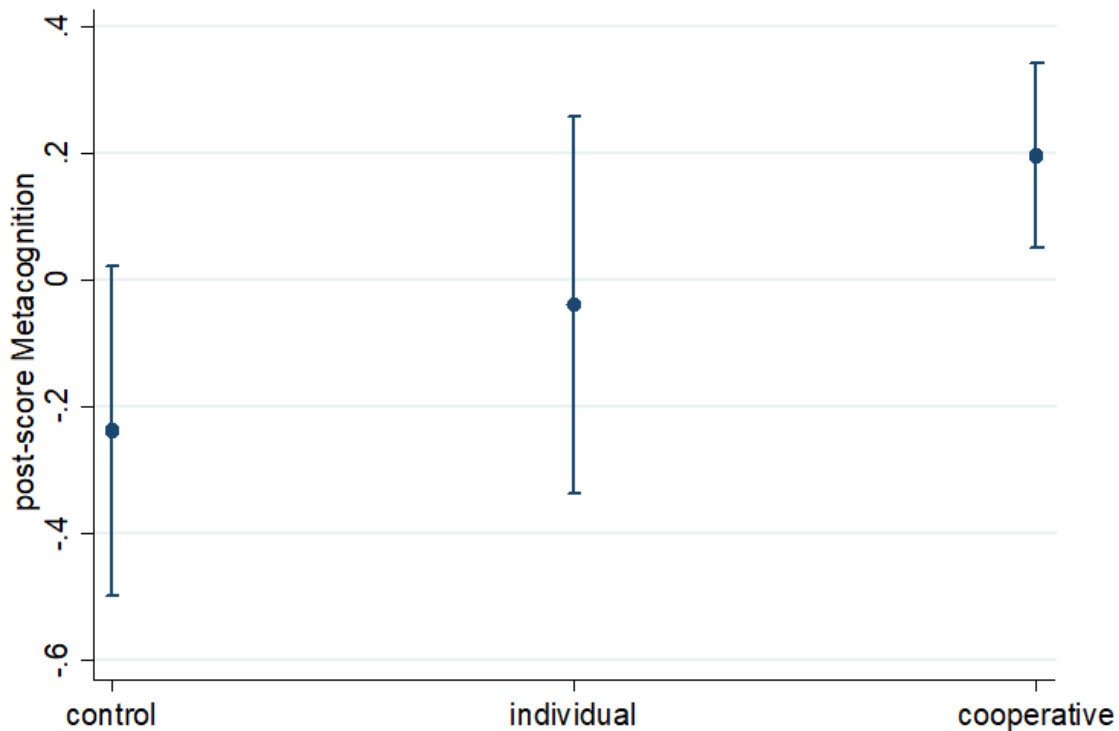
the cooperative treatment on metacognition with  $\hat{\beta}_2$  equal to 0.43 of a standard deviation, significant at the 1 % level ( $p < 0.01$ ). This implies that a combination of peer discussions and teacher feedback significantly and positively affects metacognition compared with the control condition.

A positive but not significant effect is found for the individual treatment. This is visually represented in Figure 4.4, where we see the mean standardized post-scores for metacognition as well as the confidence interval for each treatment condition. Figure 4.4 clearly shows that the confidence intervals between the control and cooperative condition do not overlap, whereas the overlap between the individual condition and the other two conditions is quite large. Figure 4.4 also shows that the mean of the cooperative condition is not only significantly different from the control condition, but is also significantly different from zero.

To improve the precision of our results, we add the pre-treatment variables metacognition, motivation, physics test score of previous year, age and *Socratic* use of previous year in Model 2, as this increases the precision of our estimates as well as the explained variance in the model, and because these variables can predict differences in outcomes (Theobald & Freeman, 2014). Adding these variables, the effect of the cooperative treatment slightly decreases, but remains significant at the 1 % level (with  $\hat{\beta}_2$  equal to 0.37 of a standard deviation).

To see whether the treatments impact metacognition of boys and girls differently, we add in Model 3 of Table 4.3 interaction terms of the treatment indicator with gender. The reference group in Model 3 are boys, which means that the effects of the treatment status  $T$  represents the effectiveness of the treatment for boys and not its effectiveness overall, as it did previously. Model 3 indicates a significant interaction effect between teacher feedback and gender ( $\hat{\theta}_1 = 0.46, p < 0.01$ ), which means that girls who receive teacher feedback score significantly higher on post-score metacognition than boys. However, we do not find significant differences for a cooperative condition between boys and girls (meaning it is effective for both genders). Furthermore, girls in general score significantly lower on post-score metacognition than boys ( $\hat{\lambda}_1 = -0.21, p < 0.01$ ), while boys with a cooperative treatment score significantly higher on post-score metacognition than boys who do not have this kind of treatment ( $\hat{\beta}_2 = 0.30, p < 0.01$ ).

Figure 4.4: Mean standardized post-scores of metacognition



- Error bars represent 95 % confidence intervals.

- The value zero is the mean of the standardized post-score metacognition of all participating students.

Recall that we divided all participating students into three subgroups of metacognitive skills (low, middle, and high metacognitive skills), based on the metacognition pre-score. Model 4 shows the effects for the students with high metacognitive skills in comparison with the effects for the other students. The coefficients of the individual and cooperative treatment therefore represent the effect for students with high metacognitive skills. Here, the interaction effect is significant for students with a cooperative treatment and a low metacognitive pre-score ( $\hat{\eta}_3 = 0.52$ ,  $p < 0.05$ ), which implies that students with low metacognitive skills experience a significantly higher effect of this treatment than students with high metacognitive skills. Furthermore, as was to be expected, students with low and middle metacognitive skills score in general significantly lower on their post measure of metacognition than students with high metacognitive skills ( $\hat{\lambda}_2 = -1.37$ ,  $p < 0.01$ ;  $\hat{\lambda}_3 = -0.43$ ,  $p < 0.05$ , respectively).

Table 4.3: Regression analyses predicting standardized post-score metacognition

	Model 1 <i>Metacognition post-score</i>	Model 2 <i>Metacognition post-score</i>	Model 3 <i>Metacognition post-score</i>	Model 4 <i>Metacognition post-score</i>
$(\beta_1)$ Individual condition ( $T_1$ )	0.20 <sup>A</sup> (0.19)	0.18 <sup>A</sup> (0.14)	- 0.031 <sup>B</sup> (0.18)	0.097 <sup>C</sup> (0.25)
$(\beta_2)$ Cooperative condition ( $T_2$ )	0.43 <sup>***A</sup> (0.15)	0.37 <sup>***A</sup> (0.12)	0.30 <sup>***B</sup> (0.12)	0.24 <sup>C</sup> (0.15)
$(\lambda_1)$ Gender (girl= 1 & boy = 0)	--	--	- 0.21 <sup>***</sup> (0.058)	- 0.0077 (0.083)
$(\theta_1)$ Individual × Gender	--	--	0.46 <sup>***</sup> (0.14)	--
$(\theta_2)$ Cooperative × Gender	--	--	0.14 (0.12)	--
$(\lambda_2)$ Metacognition Low pre-score	--	--	--	- 1.37 <sup>***</sup> (0.19)
$(\lambda_3)$ Metacognition Middle pre-score	--	--	--	- 0.43 <sup>**</sup> (0.19)
$(\eta_1)$ Indiv. × Metacog. Low pre-score	--	--	--	0.30 (0.30)
$(\eta_2)$ Indiv. × Metacog. Middle pre-score	--	--	--	0.017 (0.25)
$(\eta_3)$ Coop. × Metacog. Low pre-score	--	--	--	0.52 <sup>**</sup> (0.23)
$(\eta_4)$ Coop. × Metacog. Middle pre-score	--	--	--	- 0.048 (0.22)

Table 4.3 (continued)

	Model 1 <i>Metacognition post-score</i>	Model 2 <i>Metacognition post-score</i>	Model 3 <i>Metacognition post-score</i>	Model 4 <i>Metacognition post-score</i>
( $\gamma_1$ ) Metacognition pre-score	--	0.51*** (0.069)	0.51*** (0.071)	--
( $\gamma_2$ ) Motivation pre-score	--	0.049 (0.040)	0.041 (0.041)	0.11** (0.041)
( $\gamma_3$ ) Physics test score previous year	--	- 0.0024 (0.033)	- 0.0055 (0.034)	- 0.0043 (0.035)
( $\gamma_4$ ) Age	--	0.022 (0.050)	0.019 (0.048)	0.029 (0.052)
( $\gamma_5$ ) Socratic use previous year	--	0.12 (0.078)	- 0.0046 (0.082)	0.020 (0.079)
Constant	- 0.24** (0.13)	- 0.22** (0.11)	- 0.11*** (0.10)	0.38** (0.15)
Observations	633	633	633	633
R <sup>2</sup>	0.034	0.31	0.32	0.28

\*\*\* = significant at the 1 % level; \*\* = significant at the 5 % level; \* = significant at the 10 % level.

<sup>A</sup> Compared with the control condition.

<sup>B</sup> Effectiveness of the treatment *T* for boys.

<sup>C</sup> Effectiveness of the treatment *T* for high-metacognitive aware students.

#### 4.4.2 Results motivation

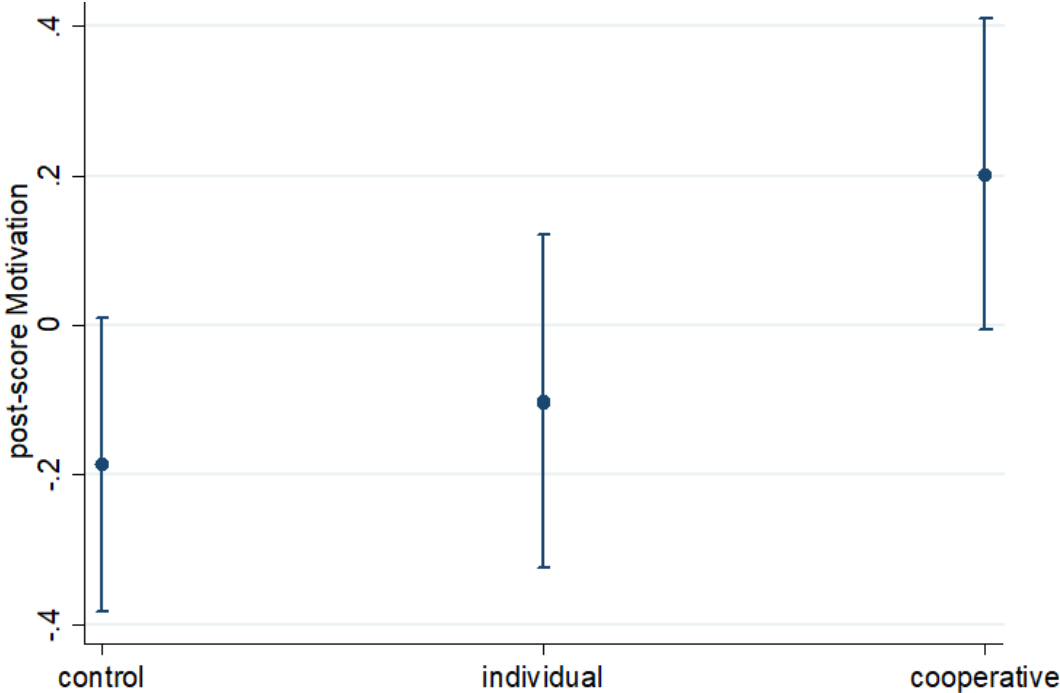
More details for the outcome motivation are presented in Models 5 and 6 in Table 4.4. Model 5 shows that the estimate  $\hat{\beta}_2$  of the cooperative treatment is equal to 0.39 points of standard deviation significant at the 1 % level ( $p < 0.01$ ). This implies that peer discussions combined with teacher feedback significantly and positively affect motivation. A non-significant effect for motivation is found for the individual treatment. This is visually represented in Figure 4.5, where we see the mean standardized post-scores for motivation as well as the confidence interval for each treatment condition.

When covariates are included in the analysis to increase the precision of our estimates, the effect of the cooperative treatment in Model 6 retained with  $\hat{\beta}_2$  equal to 0.36 point of standard deviation, significant at the 1 % level ( $p < 0.01$ ).

In Model 7, we include interactions of treatment and gender to the analysis, in order to explore the heterogeneity of the treatment effects by gender. Here, a positive significant interaction effect of treatment and gender is observed ( $\hat{\theta}_1 = 0.43$ ,  $p < 0.05$ ). This means that girls become more motivated when using an individual treatment than boys. Note that the reference group are the boys, implying that girls in general score significantly lower on post-score motivation than boys ( $\hat{\lambda}_1 = -0.29$ ,  $p < 0.10$ ), while boys with a cooperative treatment score significantly higher than boys who do not have this kind of treatment ( $\hat{\beta}_2 = 0.34$ ,  $p < 0.05$ ).

Model 8 shows the results of the effects for interaction analysis based on pre-measure metacognition of students (where students with high metacognitive skills are the reference group to students with low and middle metacognitive skills) for motivation post-score. We chose to divide groups based on pre-score metacognition instead of pre-score motivation. The reason for this is that metacognition is our main variable of interest, while motivation is a mediator and an intermediate outcome measure. Here, we do not find any significant interaction effects for students with low and middle metacognitive skills, compared with students with high metacognitive skills.

Figure 4.5: Mean standardized post-scores of motivation



- Error bars represent 95 % confidence intervals.
- The value zero is the mean of the standardized post-score motivation of all participating students.

Table 4.4: Regression analyses predicting standardized post-score motivation

	Model 5 <i>Motivation post-score</i>	Model 6 <i>Motivation post-score</i>	Model 7 <i>Motivation post-score</i>	Model 8 <i>Motivation post-score</i>
$(\beta_1)$ Individual condition ( $T_1$ )	0.085 <sup>A</sup> (0.15)	0.069 <sup>A</sup> (0.12)	- 0.13 <sup>B</sup> (0.16)	0.079 <sup>C</sup> (0.18)
$(\beta_2)$ Cooperative condition ( $T_2$ )	0.39 <sup>***A</sup> (0.14)	0.36 <sup>***A</sup> (0.11)	0.34 <sup>**B</sup> (0.15)	0.31 <sup>***C</sup> (0.11)
$(\lambda_1)$ Gender (girl= 1 & boy = 0)	--	--	- 0.29 <sup>*</sup> (0.16)	0.18 <sup>***</sup> (0.037)
$(\theta_1)$ Individual × Gender	--	--	0.43 <sup>**</sup> (0.19)	--
$(\theta_2)$ Cooperative × Gender	--	--	0.043 (0.19)	--
$(\lambda_2)$ Metacognition Low pre-score	--	--	--	- 0.16 (0.13)
$(\lambda_3)$ Metacognition Middle pre-score	--	--	--	0.018 (0.15)
$(\eta_1)$ Indiv. × Metacog. Low pre-score	--	--	--	0.010 (0.21)
$(\eta_2)$ Indiv. × Metacog. Middle pre-score	--	--	--	- 0.052 (0.24)
$(\eta_3)$ Coop. × Metacog. Low pre-score	--	--	--	0.10 (0.18)
$(\eta_4)$ Coop. × Metacog. Middle pre-score	--	--	--	0.065 (0.19)



Table 4.4 (continued)

	Model 5 <i>Motivation post-score</i>	Model 6 <i>Motivation post-score</i>	Model 7 <i>Motivation post-score</i>	Model 8 <i>Motivation post-score</i>
( $\gamma_1$ ) Metacognition pre-score	--	0.077** (0.039)	0.081** (0.039)	0.49*** (0.060)
( $\gamma_2$ ) Motivation pre-score	--	0.48*** (0.062)	0.47*** (0.062)	- 0.15 (0.083)
( $\gamma_3$ ) Physics test score previous year	--	0.18*** (0.037)	0.17*** (0.035)	- 0.019 (0.044)
( $\gamma_4$ ) Age	--	- 0.024 (0.046)	- 0.027 (0.041)	- 0.014 (0.081)
( $\gamma_5$ ) <i>Socratic</i> use previous year	--	- 0.030 (0.079)	- 0.037 (0.081)	0.49*** (0.060)
Constant	- 0.19 (0.096)	0.22 (0.76)	- 0.11 (0.16)	- 0.049 (0.11)
Observations	633	633	633	633
R <sup>2</sup>	0.030	0.38	0.40	0.39

\*\*\* = significant at the 1 % level; \*\* = significant at the 5 % level; \* = significant at the 10 % level.

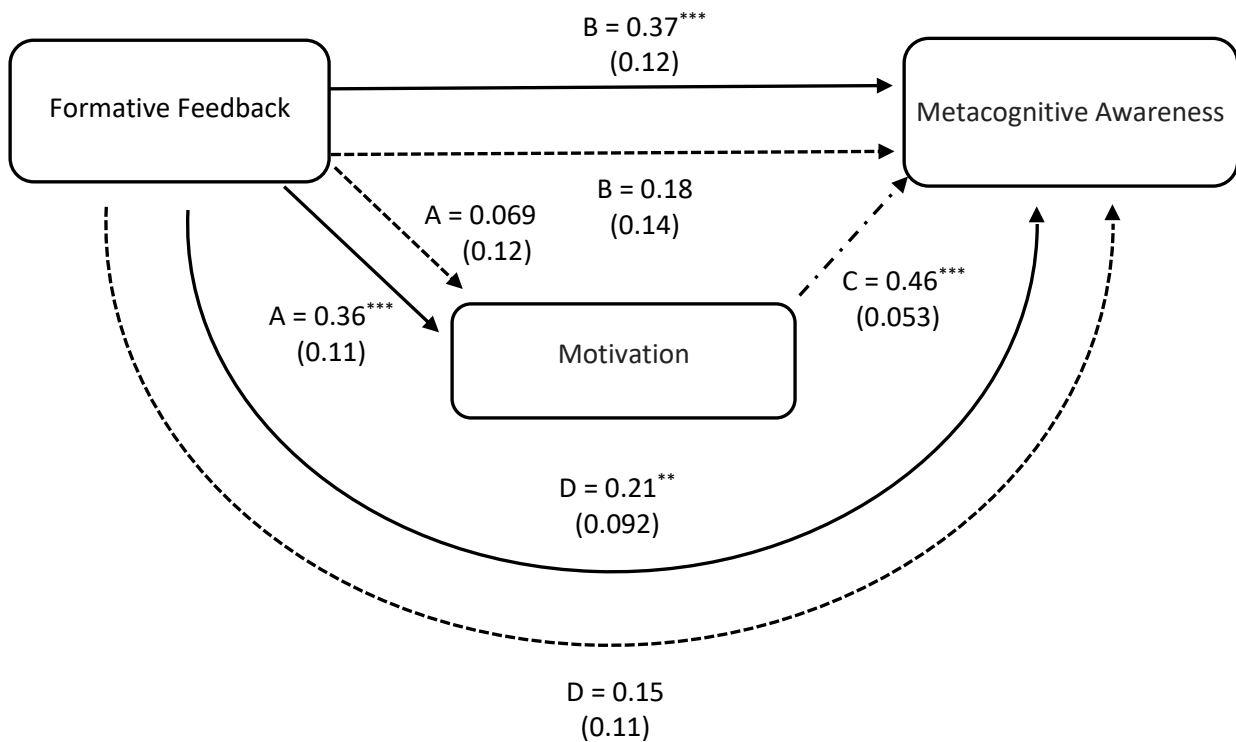
<sup>A</sup> Compared with the control condition.

<sup>B</sup> Effectiveness of the treatment *T* for boys.

<sup>C</sup> Effectiveness of the treatment *T* for high-metacognitive aware students.

#### 4.4.3 Potential mechanisms

Figure 4.6: Mediation analysis of the cooperative and individual treatment



\*\*\* = significant at the 1 % level; \*\* = significant at the 5 % level; \* = significant at the 10 % level.

- Cooperative treatment = solid arrows; Individual treatment = dashed arrows

In order to determine whether motivation mediates the estimated effects of treatment status  $T$  to metacognition, a series of four regression analyses will be conducted. The first step is to demonstrate a significant relationship between treatment status (individual or cooperative treatment) and motivation. Recall that Model 6 in Table 4.4 showed that only the cooperative treatment has a significant positive effect on motivation ( $\hat{\beta}_2 = 0.36$ ,  $p < 0.01$ ). Therefore, we focus solely on this significant effect, which is presented in Figure 4.6, path A. The second step must be conducted with treatment status predicting metacognition. Recall Model 2 in Table 4.3 showing that the cooperative treatment has a significant positive effect on metacognition ( $\hat{\beta}_2 = 0.37$ ,  $p < 0.01$ ). This effect is presented in Figure 4.6, path B. The third step includes the motivation post-score into a regression with outcome metacognition. When we do not control for treatment status, the results of Model 9 in Table 4.5 indicate that motivation significantly correlates with metacognition ( $\hat{\gamma}_6 = 0.46$ ,  $p < 0.01$ ). This correlation is presented in Figure 4.6, path C. Finally, motivation can mediate the estimated effect of treatment status and metacognition, which can be detected if the motivation post-score is

added to the regression analysis of treatment and the post measure of metacognition. Model 10 in Table 4.5 shows that the estimate  $\hat{\beta}$  of the cooperative treatment still significantly predicts metacognition ( $\hat{\beta}_2 = 0.21$ ,  $p < 0.05$ ), path D in Figure 4.6. However, the effect size is significantly reduced from 0.37 to 0.21, which suggests that motivation partly mediates the effects of the cooperative treatment on metacognition.

We also use an additional Sobel test (Sobel, 1982). According to this test, the reduction in beta coefficients between a cooperative treatment and metacognition is significant when motivation post-score is introduced into the models ( $z = 2.35$ ,  $p = 0.019$ ), confirming that motivation is a partial mediator here.

Table 4.5: Regression analyses predicting mediation analysis metacognition

	Model 9 <i>Metacognition post-score</i>	Model 10 <i>Metacognition post-score</i>
$(\beta_1)$ Individual condition ( $T_1$ )	--	0.15 <sup>A</sup> (0.11)
$(\beta_2)$ Cooperative condition ( $T_2$ )	--	0.21 <sup>**A</sup> (0.092)
$(\gamma_1)$ Metacognition pre-score	0.48 <sup>***</sup> (0.066)	0.48 <sup>***</sup> (0.067)
$(\gamma_2)$ Motivation pre-score	- 0.17 <sup>***</sup> (0.046)	- 0.16 <sup>***</sup> (0.041)
$(\gamma_3)$ Physics test score previous year	- 0.079 <sup>**</sup> (0.030)	- 0.080 <sup>**</sup> (0.032)
$(\gamma_4)$ Age	0.037 (0.039)	0.032 (0.040)
$(\gamma_5)$ Socratic use previous year	0.022 (0.077)	0.025 (0.075)
$(\gamma_6)$ Motivation post-score	0.46 <sup>***</sup> (0.053)	0.44 <sup>***</sup> (0.050)
Constant	- 0.62 (0.64)	- 0.66 (0.67)
Observations	633	633
R <sup>2</sup>	0.42	0.43

\*\*\* = significant at the 1 % level; \*\* = significant at the 5 % level; \* = significant at the 10 % level.

<sup>A</sup> Compared with the control condition.

- Standard errors are clustered at the class level.

## 4.5 Conclusion and discussion

The aim of this study was to identify the effect of formative feedback during feedback strategies with SRS on students' metacognition. Therefore, we conducted a randomized experiment among 633 students in physics classes in secondary education in the Netherlands. With respect to research questions 1 and 2, the results show that there is a significant positive effect of peer discussions combined with teacher feedback on both metacognition and motivation. The results indicate a significant standardized effect size of around 0.4 for metacognition and 0.4 for motivation. These positive findings are in line with previous literature (Blasco-Arcas et al., 2013; Buil et al., 2016; Mazur, 1997) and support the studies of Brady et al. (2013), Jones et al. (2012) and Mayer et al. (2009) that suggest that SRS usage should be implemented in conjunction with peer discussions to produce more metacognition. The meta-analysis of Hunsu et al. (2016) shows average gains of about 0.8 of standard deviation for metacognition and 0.1 of standard deviation for motivation in SRS treatments. However, these effect sizes are difficult to compare with our study, because included studies only use control groups without SRS or without questioning, whereas our control condition is formatively assessed and uses SRS. During informal discussions with students, we heard that students generally find that teacher explanations of correct reasonings are more informative and more efficient. As such, students indicate that they enjoy peer discussions; it breaks the monotony of passive listening to a teacher and teaches them to substantiate answers. Corresponding findings have also been reported by the studies of Smith et al. (2009) and Lewin, Vinson, Stetzer and Smith (2016). We assume that peer discussions stimulate students to be more active and to not just accept answers from the teacher or peer students without critical thinking. Similarly, we propose that students increase their motivation and metacognitive awareness and skills from peer discussions because we used paired sets of conceptual questions. As stated in Chapter 3, students could learn from discussions with their peers and are able to apply that understanding to the context of the follow-up isomorphic question. Discussions of pairs of questions stimulate in-depth understanding and help students to go beyond the 'plug-and-chug' or 'guess-and-check' strategy for problem solving; it develops reasoning and metacognitive skills that improves transfer of knowledge from one problem to another (Singh, 2008).

With respect to research question 3a, we find that there are differential effects among girls and boys. Compared to boys, girls score significantly higher on post-score metacognition and become more motivated when using teacher feedback. Girls increase their metacognitive skills and motivation via both an individual and a cooperative treatment, while boys increase their metacognitive skills and motivation only via a cooperative treatment. These gender effects are compatible with the previous findings of Lorenzo et al. (2006) and King and Joshi (2008) that both girls and boys benefit from feedback strategies with SRS, but that they approach learning differently. In line with previous literature, we argue that girls may have a more positive attitude to teacher feedback and are more aware about the quality of this feedback (Carvalho, Santos, Conboy & Martins, 2014; Havnes, Smith, Dysthe & Ludvigsen, 2012), which improves their perceptions of competence and metacognitive skills (Nicaise, Cogérino, Fairclough, Bois & Davis, 2007).

In answer to research question 3b, our findings show also differential effects among subgroups of students with different metacognitive skills. Students with low metacognitive skills benefit significantly more from peer discussions on top of teacher feedback than students with high metacognitive skills. This result answers the open question of Vickery et al. (2015) of whether a cooperative treatment benefits students with low metacognitive skills more than students with high metacognitive skills. A possible explanation for our findings may be that students with low metacognitive skills highly benefit from the interaction with high-metacognitive peers. Additional analyses indeed shows that the level of metacognition of peers in the classroom positively relates to an individual's level of metacognition, as can be seen in Table 4.7 in Appendix 4.3. This is in line with the conclusions of Schraw, Olafson, Weibel and Sewing (2012) and Shapiro et al. (2017) who show that students with low metacognitive knowledge and skills benefit from instruction and collaborating with a more experienced, metacognitive learner. Our findings are also in line with the findings of Shute (2008), that formative feedback reduces the cognitive load of students during learning, especially students with low metacognitive skills with more complex problem-solving questions who can be cognitively overwhelmed due to high performance demands. A reason for our results may be that multiple-choice questions with an 'application' or 'analyzing' level have more meaning for students with low metacognitive skills than students with high metacognitive skills, because these questions require less high metacognitive skills. However, the studies of Knight

et al. (2013) and Knight, Wise, Rentsch and Furtak (2015) show that more cognitively demanding multiple-choice questions with a 'synthesis' or 'evaluation' level do not necessarily benefit students with high metacognitive skills; students in general are significantly more engaged to discuss lower-order Bloom's level questions, because the argumentations needed to support the answers are less complex.

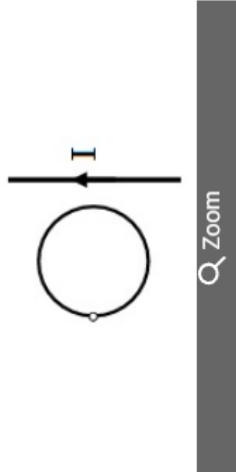
Finally, in response to research question 4, we show that the effect of peer discussions combined with teacher feedback on metacognition runs partly through its influences on motivation. This means that assessments with peer discussions, on top of teacher feedback, significantly motivate students, which in turn partially affects metacognition. This is a novel finding and contribution to the literature, as no prior study documented that motivation partly mediates the effects of a cooperative treatment and metacognition. We argue that peer discussions encourage monitoring and develop skills for determining how well students understand course material, which directly improves metacognition. Similarly, peer discussions make formative assessments more interactive and interesting, and make students more proactive and focused when answering, which leads to more motivation to learn and, indirectly, also to more metacognitive awareness and skills. One could argue that the improvements of metacognitive skills is attributed to a larger amount of time spent on these questions, during the peer discussions. However, each assessment lasted about the same time, regardless of the condition, so there was no extra time.

With respect to the generalization of the results, we are convinced that the results also apply to most secondary schools in the Netherlands, because our six schools are representative of the average secondary school in the Netherlands. All statistics for these schools are within half a standard deviation of the average of all variables. Furthermore, we used the cloud-based student response system *Socrative*, that has similar options and specifications (e.g. plotting histograms) as other commonly used audience response systems and would therefore very likely lead to the same results. However, the generalization of the results is limited to countries in which all students have the possession of a mobile phone or other mobile device, and are used to working with educational technology in class. Also further evidence using data from other educational contexts should establish how generalizable the results are. Finally, we should note that the results in this chapter are based on self-reported questionnaires, which prompt students to recall events and align their responses to the question's response

scales. Although these questionnaires are easy to administer (and thus valuable for practice and large-scale use), they have the disadvantage that students may interpret questions differently or fill in socially desirable answers, which may result in biases (Harrison & Vallin, 2018; Schellings & Van Hout-Wolters, 2011). However, the study of Harrison and Vallin (2018) showed that if any bias exists, it is consistent between conditions, which means that statistical tests between conditions are not likely influenced by such biases.

## Appendix 4.1

Figure 4.7: Two screenshots of a paired question in *Socrative* on smartphones screens



**Question A**  
A current-carrying wire lies on the ground.  
A circular coil is placed next to the wire.  
The current in the wire suddenly grows stronger.  
What is the direction of the induced current in the coil?

A The induced current changes its direction from clockwise to counter-clockwise.

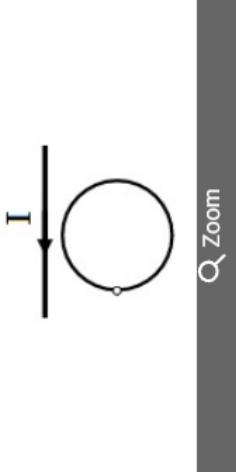
B Clockwise

C Counter-clockwise

D There is no induced current in the coil

E I don't know

SUBMIT ANSWER



**Question B**  
A current-carrying wire lies on the ground.  
A circular coil is placed next to the wire.  
The current in the wire vanishes suddenly.  
What is the direction of the induced current in the coil?

A The induced current changes its direction from counter-clockwise to clockwise.

B Clockwise

C Counter-clockwise

D There is no induced current in the coil

E I don't know

SUBMIT ANSWER



## Appendix 4.2

Table 4.6: T-tests between included and excluded students

	<i>Included students</i> ( <i>N</i> = 633)		<i>Excluded students</i> ( <i>N</i> = 108)		<i>t</i>	<i>p</i>
	<i>Mean</i>	<i>Std. Dev.</i>	<i>Mean</i>	<i>St. Dev.</i>		
Gender <sup>A</sup>	0.47	0.50	0.56	0.50	- 1.62	0.11
Age	15.73	1.03	15.86	0.77	- 1.25	0.21
Physics score <sup>B</sup>	6.55	1.00	6.41	1.09	1.27	0.20
<i>Socratic</i> use <sup>B</sup>	0.78	0.42	0.76	0.43	0.38	0.71
Metacognition <sup>C</sup>	3.28	0.37	3.23	0.40	1.23	0.22
Motivation <sup>C</sup>	3.44	0.54	3.39	0.52	0.92	0.36

\*\*\* = significant at the 1 % level; \*\* = significant at the 5 % level; \* = significant at the 10 % level.

<sup>A</sup> boy = 0, girl = 1

<sup>B</sup> previous year

<sup>C</sup> average pre-score

## Appendix 4.3

Table 4.7: Regression analysis post-score metacognition

	<i>Metacognition post-score</i>
Metacognition pre-score	0.41*** (0.080)
Motivation pre-score	0.081 (0.051)
Gender (boy = 0 & girl = 1)	- 0.061 (0.097)
Physics test score prev. year	- 0.028 (0.037)
Age	- 0.025 (0.057)
<i>Socratic</i> use prev. year	0.077 (0.13)
<b>Metacognition pre-score PEER student</b>	<b>0.24***</b> <b>(0.058)</b>
Constant	0.13 (0.11)
Observations	264
R <sup>2</sup>	0.29

\*\*\* = significant at the 1 % level; \*\* = significant at the 5 % level; \* = significant at the 10 % level.

- ONLY for students in the cooperative condition.

- Standard errors are clustered at the class level.



## Chapter 5

### A Conceptual Framework to Understand Learning:

### The Role of Prompts and Diagnostic Cues <sup>18</sup>

---

<sup>18</sup> This study is based on joint work with Anique de Bruin and Carla Haelermans.

The aim of this chapter is to develop a conceptual framework to identify relationships between factors that may be responsible for influencing metacognition when students are formatively assessed using educational technology such as student response systems. We describe and bridge the monitor and control model of Nelson and Narens (1990) and the cue utilization framework of Koriat (1997), which we complement with aspects from digital formative assessments. Based on this we develop a conceptual framework that provides insight into which prompts may influence the utilization of diagnostic cues and thereby learning. The developed framework suggests that more prompts lead to more diagnostic cues, which improve students' accuracy of monitoring judgments and enhanced metacognition.

## 5.1 Introduction

The construction of a learning environment in which providing feedback on performance is intended to improve learning is important in education and it is a priority of teachers. This is often realized by formative assessments, which according to Black and Wiliam (1998a) is a general term for all those activities undertaken by teachers, which provide information and are used as feedback to modify teaching and learning activities (p. 7). Educational technology, such as ‘polling systems’ or ‘student response systems’ (SRS), is increasingly used for this (Edens, 2008; Smith et al., 2011). These systems are electronic voting devices (e.g. clickers) or voting software that support mobile devices (e.g. *Socrative* and *Kahoot!*), allowing teachers to obtain real-time student performance information via (anonymous) answers to various types of questions (mostly multiple-choice questions). The formative character of the assessments here is exhibited mainly in the way in which the activity creates opportunities for students to share their thinking with their teacher and peers without grading (Bennett, 2011; Krumsvik & Ludvigsen, 2012; Ludvigsen et al., 2015; Mazur, 1997).

During the last two decades, a great deal of research has been conducted into formative assessments with SRS, with a particular focus on students’ understanding (e.g. Egelandstal & Krumsvik, 2017b; Krumsvik, 2012; Ludvigsen et al., 2015), achievement (e.g. Cohn & Fraser, 2016; Hwang & Chang, 2011; Molin, Cabus, Haelermans & Groot, 2019), active engagement (e.g. Carnaghan & Webb, 2007; Kay & LeSage, 2009; Lantz & Stawiski, 2010) and participation (e.g. Han & Finkelstein, 2013; Levesque, 2011; Oigara & Keengwe, 2013). However, little attention has been paid to the extent to which these assessments influence metacognition (Brady et al., 2013b), resulting in a lack of understanding of how and when SRS works best (Brady, Rosenthal, Forest & Hocevar, 2020). Because formative assessments focus on cognition itself, the prefix ‘*meta*’ is added here to indicate that metacognition is about or above cognition. Cognition refers to processes such as problem-solving and memory, while metacognition refers to the awareness of students’ own thinking; it is thinking about and controlling one's own thinking and learning (Schunk & Zimmerman, 1998). Some researchers suggest that formative assessments with SRS encourage students to engage in metacognitive processes when students become aware of their own understanding relative to that of their peers (Anthis, 2011; Brady et al., 2013a; James & Willoughby, 2011; Mayer et al., 2009; Noel,

2010; Tullis & Goldstone, 2020). In Chapter 4, we conducted an empirical study of students' metacognitive awareness in classes where formative assessments with SRS are conducted. We showed that peer discussions combined with teacher feedback significantly positively affect metacognition. The studies of Brady et al. (2013b) and Brady et al. (2020) both demonstrated that metacognition during assessments with clickers has a more productive influence (i.e. more self-reflection with the least possible interference from, for example, uncomfortable social comparisons) on the learning process than assessments with paddles (a type of flashcard system). Jones et al. (2012) showed that students increase their metacognitive awareness via SRS-based instruction. All studies did not, however, conceptually describe how formative assessments with SRS may help learning and which relationships between students' behavior and thought processes might be responsible for affecting metacognition.

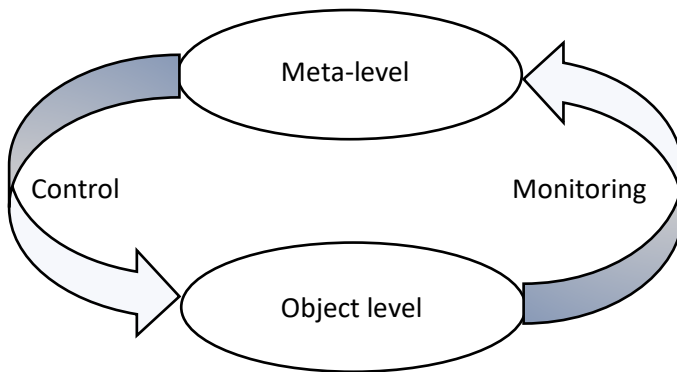
Many teachers use SRS in their formative assessments without realizing the extent to which the questioning method may affect student learning, i.e. how students judge their learning and make follow-up study decisions based on that assessment. Given the rise of SRS, it is important to identify best-practices for how to implement formative assessments in order to improve metacognitive accuracy and, ultimately, learning. This study attempts to develop a conceptual framework that represents the relationships between factors that may be responsible for affecting metacognitive awareness when students are formatively assessed using SRS. We will first introduce the monitor and control model of Nelson and Narens (1990) and describe how the cue-utilization framework of Koriat (1997) can be used as a basis for monitoring judgments. Based on this framework, we develop a conceptual framework with the main aim being to shed more light on how prompts may enhance the utilization of diagnostic cues and thereby increase learning. Finally, we present the main conclusions and sketch directions for future research.

## 5.2 Theory on metacognition: monitoring and control

The term metacognition has been used in many different ways in previous literature and theory, but an important common denominator is that metacognition involves monitoring and control (or regulation) of cognitive processes (Biggs, 1985; Butler & Winne, 1995; Pintrich, 2000a; Winne & Hadwin, 2008; Zimmerman, 1989; Zimmerman, 2000). For this reason, the

starting point of our conceptual framework is the model of Nelson and Narens (1990), in which the central role of monitoring and control of learning is depicted in a straightforward way.

Figure 5.1: Relation between monitoring and control



- Nelson and Narens (1990, p. 126)

Nelson and Narens (1990) developed a two-level model in which they established the cyclical, hierarchical relationship between cognition and metacognition, and conceptualized cognitive processes by distinguishing two levels: the *object level* and the *meta-level* (see Figure 5.1). Cognitions at the object level include task-relevant knowledge and a repertoire of automated strategies that facilitate learning and remembering. For example, students solving a mathematical cube problem like: *“Someone is building a 2 by 2 by 2 cube with blocks, and wants to build a new cube that is four times as large. How many blocks does this person need for this?”*. In this example, cognition at the object level is that students know that the ratio of the blocks between the small and the large cube is equal to  $L^3 = 4^3 = 64$ , so that a four times as large cube has 256 blocks. Cognition at the meta-level contains models of a task and cognitive operations that are necessary to perform that task, e.g. when students realize that they do not know the correct rules to solve this cube problem. The two cognition levels are connected by monitoring and control signals (Figure 5.1). During learning, information flows continuously between the two levels. The meta-level monitors cognition and thoughts at the object level. Such monitoring may include, for example, students’ judgment of their own confidence when solving a problem, the speed with which a solution strategy comes to mind, or the time required to complete a partial solution step. The degree or accuracy of monitoring determines how well information studied about the problem will be recallable in the future. At the meta-level, students compare their learning with the desired degree of learning. They do not only decide what information should be studied, but also how and when it should



be studied. Students then control or regulate their learning at the object level (Figure 5.1), by initiating, altering or terminating mental and physical actions (e.g. when repeating the same learning task or moving on to the next). In the cube problem above, students may realize that they are insufficiently familiar with the square-cube rule when solving geometric problems in their mathematics book. They may then decide to restudy their notes about the square-cube rule and/or make similar example assignments in their book, with the assumption that they will be more familiar with this rule after doing so. According to Nelson and Narens (1990), this cycle is repeated until students' learning goals are achieved.

The Nelson and Narens framework describes learning as an iterative process, where monitoring informs control actions and control actions influence learning (Nelson & Narens, 1990). As stated above, such monitoring-control processes are present in most frameworks of metacognition and they emphasize the important role of the quality of monitoring judgments that underlies regulation of learning and performance. As a next step in developing a conceptual framework, it is important to understand which factors have an impact on accuracy of monitoring judgments, i.e. the degree to which students are able to distinguish between well and less well-learned course material (Engelen, Camp, van de Pol & de Bruin, 2018).

### 5.3 Diagnostic cues to monitor learning

When students judge their learning, i.e. when making a monitoring judgment, this judgment is based on whatever information is most available at that time. This information is usually referred to as 'cues'. It refers to *"any bit of information that might potentially be drawn upon or referred to by a [...] to inform a judgment"* (Snow (1968), as cited in Cooksey, Freebody & Wyatt-Smith, 2007, p. 431). Using cues to inform monitoring judgments about the state of learning is referred to as cue-utilization (Koriat, 1997). Koriat's cue-utilization framework states that cues that lead to accurate monitoring judgments are those that are predictive of subsequent test performance. Since students do not have direct access to the quality of their learning processes or their performance, they need to infer cues based on, e.g. task information, personal information, and contextual information. For example, students' study time in self-paced study or the subjective ease with which information is processed or retrieved are cues that could inform students' judgments when assessing their learning and

testing performance (Benjamin, Bjork & Schwartz, 1998; Hoffmann-Biencourt, Lockl, Schneider, Ackerman & Koriat, 2010; Kelley & Rhodes, 2002; Thiede & Dunlosky, 1999). These cues are indicative of processing fluency, as more time needed for studying or retrieval is suggestive of greater difficulty and potentially indicates poorer test performance (Efklides, Schwartz & Brown, 2018). The feeling of difficulty or the feeling of familiarity in recognizing a solution strategy are other cues that are used by students to judge their learning. If students are not aware of the correctness of an answer, they have to rely on these cues when, for example, they are formatively assessed. (We must note that familiarity with a topic can be misleading to use as a cue as it tends to be a poor predictor of test performance.)

The examples described indicate the difficulty students experience to accurately monitor their learning. Cues are only predictors of subsequent learning and test performance when they accurately reflect the mental representation of the learning material that will be assessed at testing (Thiede, Griffin, Wiley & Anderson, 2010; Thiede, Redford, Wiley & Griffin, 2017). For example, the extent to which accurate solution strategies come to mind in a limited amount of time is an important prerequisite for summative assessments, and therefore a valid cue for predicting test performance during a formative assessment. But simply recognizing multiple-choice questions from previous formative assessments, is an example of an invalid cue for predicting test performance, because actual understanding of course content depends on students' ability to recognize and apply solution strategies, and not only on recognizing a question from earlier assessments. The use of (or reliance on) non-diagnostic and invalid cues leads to inaccurate monitoring judgments and has negative implications for the effectiveness of control decisions. To increase access to diagnostic cues, formative assessments should be well aligned with summative assessments to predict test performances on summative assessments (Stiggins & Chappuis, 2006; Thiede et al., 2017).

## 5.4 A conceptual framework

In previous studies, formative assessments have been described as frequent, interactive assessments that identify students' progress and learning needs to adapt teaching appropriately (Looney, 2005). Formative assessments are those activities undertaken by teachers and students that are intended to provide feedback to improve student test performance (Black & William, 1998a; Nicol & Macfarlane-Dick, 2006; Sadler, 1998). One approach to provide feedback in daily lessons are formative assessments using SRS. Using these types of assessments create opportunities for students to regularly share their thinking with teachers and peers (Ludvigsen, Krumsvik & Breivik, 2020; Smith et al., 2011) with limited additional workload for the teacher during the assessment (de Sousa, 2018; Nicol & Macfarlane-Dick, 2006). A deeper understanding of students' behavior and thought processes when using digital formative assessments with SRS throws light upon which relationships between these aspects affect metacognition. By means of a conceptual framework (Figure 5.2), we discuss which prompts in formative assessments with SRS can influence the utilization of diagnostic cues, which in turn affect learning.

### 5.4.1 Prompts and diagnostic cues

When teachers use formative assessments and pose multiple-choice questions to students, they are checking students' understanding by using prompts. Here, *prompts* are defined as activators or activities which help students to engage in their cognitive processes. Prompts trigger previous knowledge of students and are expected to identify knowledge gaps and understanding problems that activate students to improve their accuracy of monitoring judgments, such as anonymous polling prompts, visual feedback prompts (e.g. histograms, check marks), teacher feedback prompts, peer discussions prompts, and isomorphic question prompts. See Figure 5.2. Based on the definition of prompts as activators of cognitive processes, *cues* are defined as information that shift students attention to the required knowledge and that are diagnostic for cognitive processes and test performance. Prompts are needed to improve the use of diagnostic cues when judging monitoring and regulation of learning (De Bruin & Van Merriënboer, 2017; Van Loon, de Bruin, van Gog, van Merriënboer & Dunlosky, 2014). The key to improving monitoring judgments lies in providing students with diagnostic cues that predict performance on summative assessments. Such cues are theory-

based cues and experience-based cues (Koriat et al., 2008). Theory-based cues (also called knowledge-based cues; Bjork, Dunlosky & Kornell, 2013) refer to what students consciously believe about their knowledge. These cues are related to the beliefs students might have about their competence. When knowledge is unconsciously acquired or processed, theory-based cues may also influence experience-based cues (Koriat et al., 2008), as cues influence one another (Roelle, Nowitzki & Berthold, 2017). Experience-based cues are derived from emotions or ‘sheer subjective’ feelings and include anything students directly experience during learning (Koriat, 1997). Emotions are crucial and connect to metacognition, because through sheer subjective feelings students are aware of their beliefs and ideas about cognition (Efklides, 2008). Vice versa, experience-based cues may serve as a basis for theory-based cues, as emotions are predictive of awareness of (not) comprehending course content (Koriat et al., 2008).

## 5.5 Diagnostic cues in formative assessments with SRS

In this section, we describe four cues that may be responsible for affecting metacognitive awareness when students are formatively assessed with SRS. These cues are primary and useful cues that arise from polling. They can be ‘easily’ influenced by a teacher and their effects have already been partially examined in Chapters 2 through 4.<sup>19</sup>

### 5.5.1 Comprehension cues

Ideally, prompts improve the accuracy of monitoring judgments with *comprehension cues* (Figure 5.2; Koriat, 1997), as these cues are highly valid predictors of subsequent test performance (Griffin, Jee & Wiley, 2009; Wiley, Griffin & Thiede, 2005). We define comprehension cues as information that indicates the knowledge and the quality of cognitive states so that it can be used in new contexts. These cues indicate comprehension of tasks and topics (Van Merriënboer & De Bruin, 2019). Examples of comprehension cues are (1) the amount of clues and information that is actually recognized when students read a multiple-choice question or discuss it with their peers, or (2) the number of multiple-choice questions correctly answered in an assessment. A theory-based mechanism here would be if students

---

<sup>19</sup> Note that Section 5.3 refers to fluency. Fluency may also be one of the cues in the framework of Figure 5.2. However, we choose to use cues that can be easily influenced by the teacher, but we are also aware that other cues (not included in this framework) can also be influenced by the teacher.

reason that if they earn low scores on the polls, they will likely earn a low score on the exam, too. In contrast, an experience-based mechanism would be that the poll felt difficult, they think the exam will also be difficult. Students who are aware of comprehension cues experience increased understanding of the course material and can apply metacognitive skills.<sup>20</sup> They experience that they are competent in their learning and problem solving, how to coordinate it and how to monitor it. However, the way in which an assessment is implemented can dramatically alter whether comprehension cues are diagnostic of learning or not. For example, if an assessment is done immediately after learning new course content, and not after a delay, the comprehension cues will be far less diagnostic (e.g. Myers, Rhodes & Hausman, 2020; Rhodes & Tauber, 2011).

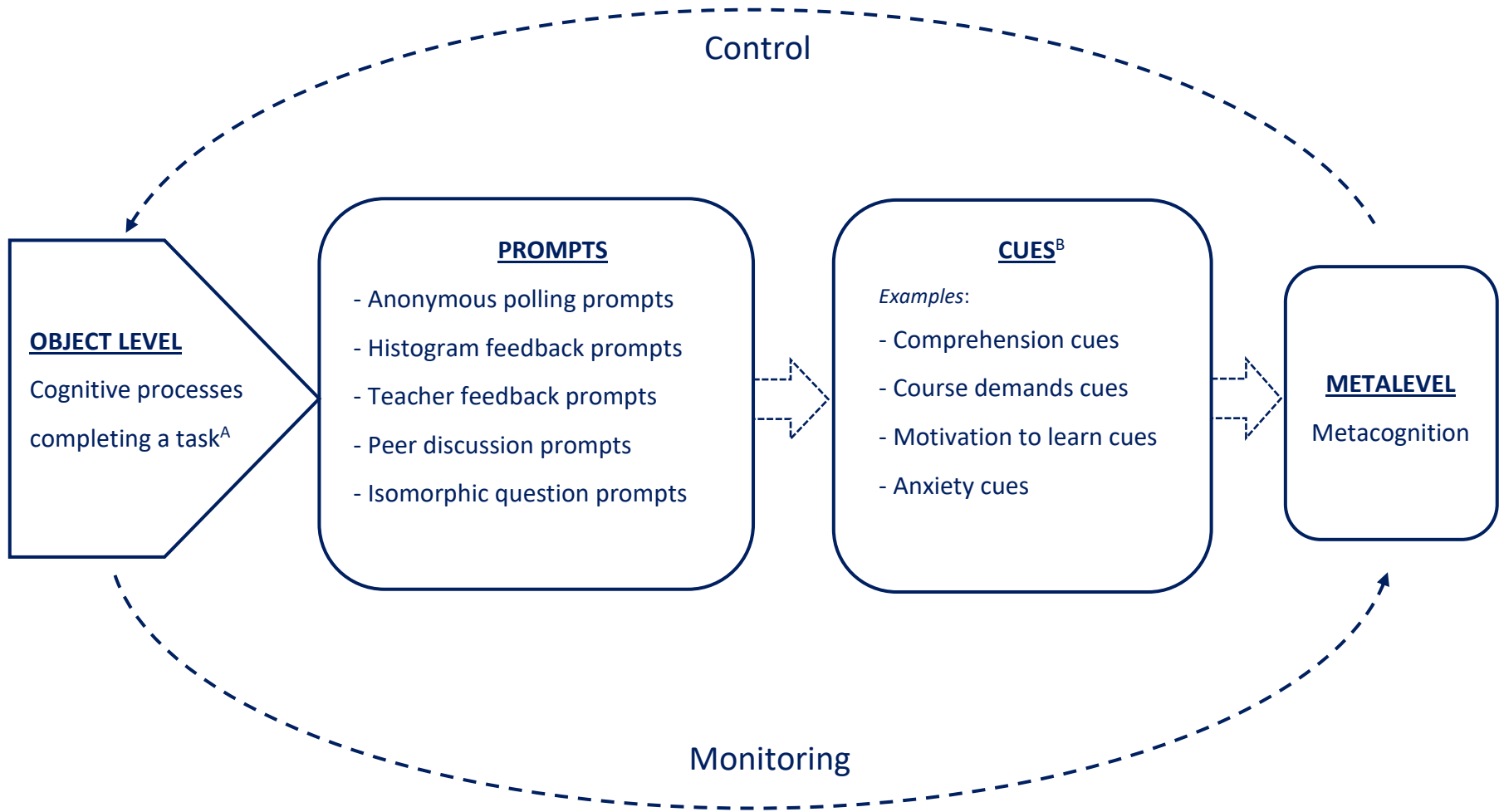
### 5.5.2 Course demands cues

Another type of cues are *course demands cues* (Figure 5.2), which we define as information that make students aware of what is expected during course. Through prompts in formative assessments, students can meet the cognitive demands of summative assessments and improve their meta-level by recognizing information about, for example, the type of posed questions, the level of abstraction, the difficulty of the questions, and the skills or solution strategies needed to answer the questions (Isaacson & Fujita, 2006; Maule, 2001). It is primarily the repetitive nature of certain questions and the teacher's emphasis on solution strategies that inform students of the course content and the importance of questions in summative assessments (Efklides, 2006; Mitchum & Kelley, 2010). Being aware of the emphasis on these questions enable students to develop metacognitive skills and knowledge, because adjusting solution strategies such that they are in line with the demands of the course (Ross, Green, Salisbury-Glennon & Tollefson, 2006; Winne & Hadwin, 1998) or adapting the study-behavior and study-planning based on information on what teachers expect to see in an exam are metacognitive processes (Entwistle & Entwistle, 2003).

---

<sup>20</sup> Metacognitive skills are, for example, that after answering a multiple-choice question students know how to easily organize gained information, wonder if they answered the question correctly, consider if there is not an easier way to answer the question, etc.

Figure 5.2: A developed framework of prompts and cues by formative assessments with SRS



<sup>A</sup> The SRS and the prompts in the formative assessments are the context in which a task is completed.

<sup>B</sup> Theory-based cues and experience-based cues

### 5.5.3 Motivation to learn cues

Another cue for predicting test performance is, for example, *motivation to learn* (Figure 5.2; Efklides et al., 2018). When trying to understand a problem, not only are cognitive processes important, but also motivational processes. According to Ryan and Deci (2000), motivation to learn may be defined as ‘to be moved to learn’. Motivation plays an important role in the monitor and control model of Nelson and Narens (1990), as exercising control is a strenuous process that requires motivation (Efklides, 2011). It provides the basis for judgments and actions (Koriat & Levy-Sadot, 2000). During learning, students constantly evaluate how much pleasure they have while performing a task and how much confidence they have in finding a correct solution. This pleasure and confidence is related to what students experience during learning and is linked to metacognitive experiences. Students who have more pleasure and confidence during learning are more motivated to learn, and are more willing to choose solution strategies in new situations than students who lack motivation (Pekrun, 2006; Pintrich & De Groot, 1990). Motivated students may use motivation as a cue and judge themselves higher on a formative or summative assessment. They open themselves up to problems they have not solved before, while actively searching in their memory for relevant strategies, content, and task knowledge (Schunk & Zimmermann, 2009). Motivated students set goals for themselves and put in more effort into succeeding. This allows students to use metacognitive strategies that enable them to learn and understand deeply (Bandura, 1993; Sungur & Senler, 2009). Motivation is often considered a prerequisite for monitoring and regulation (Zimmerman, 2000). The study of Baars and Wijnia (2018) demonstrated that motivated students score higher on monitoring accuracy and learning outcomes than unmotivated students. On the other hand, gaining metacognitive knowledge of oneself as a learner also contributes to viewing oneself as a competent learner, which not only affects success in learning, but also motivation to learn (White & Frederiksen, 2005; Zimmerman, 1989).

### 5.5.4 Anxiety cues

Another example of a cue that predicts test performance and affects metacognition is *anxiety* (Figure 5.2; Elliot & McGregor, 1999; Herman, 1990; Morsanyi, Cheallaigh & Ackerman, 2019). Anxiety is an emotion that can be so overwhelming that even students with high intelligence or students who worked diligently cannot perform well. Anxiety arises when students view an

assessment as threatening, and believe that their skills are lacking for the demands of the assessment (González et al., 2017; Pekrun, 1992; Spielberger & Vagg, 1995). Although some degree of anxiety might be helpful in the learning process (Hong, 2010), a (too) high degree of anxiety might lead to an absorption of information processing, resulting in a narrowly focused processing of information that obstructs creativity in solving problems (Pekrun, 1992). Anxiety results in a reduced flexibility in attempting problem solving strategies, and, due to a tendency to avoid it, a greater chance of ending problem-solving strategies too early, that hinders monitoring and control processes (Morsanyi et al., 2019). Only in the case of relatively easy tasks, where the likelihood of success is plausible, anxious students will use more resources to achieve a high level of confidence than less anxious students (Morsanyi et al., 2019). Previous studies showed a negative relationship between anxiety and metacognition (e.g. Karasel, Ayda & Tezer, 2010; Veenman, Kerseboom & Imthorn, 2000). As anxiety prevents students from using problem-solving strategies properly, it disrupts metacognitive processes (Tobias & Everson, 1997; Veenman et al., 2000). Even irrelevant thoughts, caused by anxiety, may hinder students' metacognitive processes, which finally results in a reduction of test performance (Birenbaum & Nasser, 1994; Everson, Smodlake, Tobias, 1994; Miesner & Maki, 2007; Tobias, 1985). Students experiencing anxiety may use anxiety as a cue and judge themselves to perform poorly on a formative or summative assessment. The study of Lusk (1981) showed that students with high anxiety gave lower predictions when estimating their performance after an exam than students with low anxiety. Lusk (1981) reported that highly anxious students made more accurate predictions about their performance as they reduce their predictions, which lowers their level of overconfidence in comparison with low-anxious students. Given that most students are overconfident in their performance, Morsanyi et al. (2019) argued that students with high anxiety (due to their low levels of self-confidence) may better calibrate their metacognitive judgments, as they may have weaker tendencies toward overconfidence. The empirical studies of Miesner and Maki (2007) and Dunlosky, Rawson and Hacker (2002) both showed that anxious students use anxiety cues about how well they understood course material. That is, they are more able to judge their levels of anxiety during an assessment, and this may help them to judge their future test performance more accurately.



## 5.6 Prompts in formative assessments with SRS

To help students in identifying diagnostic cues, carefully designed interventions, such as prompts, are a requirement (De Bruin et al., 2017). Prompts enhance the use of diagnostic cues and should be inserted (to ensure reliable use) just before monitoring and controlling learning (De Bruin et al., 2017). In this section, we describe five prompts (see also Figure 5.2) that may influence monitoring judgments when students are formatively assessed with SRS. The five prompts are common components of implementing a formative assessment poll in a classroom. Previous literature showed that these elements in the framework can impact metacognition. The prompts are all individually relevant activities in an assessment, but might lead to even more effective cues when used jointly, as is explained below.

### 5.6.1 Anonymous polling

SRS allow students to respond to multiple-choice questions presented in front of the class. By using clickers, *Socrative* or *Kahoot!*, all students can simultaneously transmit their answers anonymously, after which automatically-generated graphs illustrate the distribution of answers. Anonymous polling is the first prompt (Figure 5.2). The intention of voting anonymously is to stimulate students' willingness to participate and to help them to relax while answering questions (Zhonggen, 2017). Especially for shy or anxious students, the anonymity of SRS offers an opportunity to engage in an activity that protects them from the anxiety of being noticed by others (Brady et al., 2013a; Ulbig & Notman, 2012). The anonymity releases students' anxiety and nervousness (Zhonggen, 2017). It facilitates students with *anxiety cues* by making them more confident in finding the correct solution and aware of their own cognitive skills. This is proven by the empirical study of Vollmeyer and Rheinberg (1999), who showed that less anxious students are more confident, and choose more strategies and gain more understanding when solving problems. Ulbig and Notman (2012) demonstrated that the anonymity helps shy students gain more comprehension of the course content than when traditional classroom teaching methods are used. Moreover, as compared to other assessment methods where the answers given are visible to students (e.g. raising hands, raising coloured flashcards, or applause), the anonymity eliminates the effect of students waiting for others before deciding their answer (Brady et al., 2013a). A consequence may be that anxious students experience a more stringent monitoring and

control processes by, for example, spending longer time on problems or double-checking answers before submitting them (Morsanyi et al., 2019). By solving problems on their own, students use more resources, which fosters confidence in their own intellectual abilities to answer problems correctly and improves their metacognitive awareness (Fies & Marshall, 2006; Richardson, Dunn, McDonald & Oprescu, 2015; Ulbig & Notman, 2012; Veenman et al., 2000).

### 5.6.2 Histogram feedback

When students answer multiple-choice questions with their SRS, voting software automatically aggregates class responses and presents the given answers in visuals such as histograms. These visuals show the correct answer indicated with a check mark (or with another colour) and offer real-time feedback to students as to how well concepts are being understood. Histogram feedback is the second prompt students can receive (Figure 5.2). A histogram allows students to make social, normative comparisons relative to their peers, which can serve as diagnostic cues about their own learning. Showing voting results provides students with *comprehension cues*. When students answer a question incorrectly, seeing the correct and incorrect answer options may give students the insight that they do not comprehend the course content and that they have to change their preparation or learning strategy (Chien et al., 2016; King & Joshi, 2008; Sun, 2014). The histogram can cause students to face their lack of preparation or problem solving skills and make them more metacognitively aware of their failure to succeed on a subsequent test (Covington, 1985; Naveh-Benjamin, 1991). A confirmation of comprehension is experienced when students answer the question correctly, while the majority of the class answers the question incorrectly (Kay & LeSage, 2009). However, a histogram can also lead to non-diagnostic comprehension cues. This may occur when it is not clear what the correct answer is because, for example, no check mark is shown in front of the correct answer. Comprehension cues are then non-diagnostic, because the metacognitive judgments are based on social comparisons, rather than objective norms. Social comparisons are less diagnostic with respect to whether or not a question will be answered correctly on a subsequent assessment. For this reason, it is important that the histogram shows the correct answer, and that it is followed by other prompts, such as teacher feedback, peer discussions or isomorphic questions (Brady et al., 2013a; Chien et al., 2016; Knight et al., 2013). Histograms may also help students to eliminate the feelings of

hopelessness or fear of failure, especially for anxious students who answer a question correctly. These students experience the feeling that they possess knowledge that others do not, which decreases the feelings of fear of failure (Ulbig & Notman, 2012). When giving an incorrect answer, anxious students may realize that they are not alone in struggling with course content, which reduces self-doubt and the feeling that they are unable to comprehend the content (Chien et al., 2016; Hoekstra & Mollborn, 2012; Kay & LeSage, 2009).

### 5.6.3 Teacher feedback

Showing voting results is not only a source of information for students, but also for teachers. In the most basic way of formative assessments with SRS, teachers give immediate content feedback, based on the distribution of answers, by describing the thought processes required to arrive at the correct answer (Dori, Mevarech & Baker, 2018). Usually, the feedback here is primarily a monologue or a 'one-way transmission of information from teachers to students' (Boud & Molloy, 2013, p. 702) and aimed at providing students clarity by identifying gaps in knowledge and helping them to correct misunderstandings and flaws in logic (DeBourgh, 2008). Teacher feedback is clear and accurate and serves as a third prompt (Figure 5.2). Teachers have subject-matter knowledge, use subject-specific terms and provide content-related feedback focused on learning goals (Lee, Keh & Magill, 1993; Caldwell, 2007). They may give hints, solve problems with a step-by-step walkthrough and demonstrate how to write this down in an exam. Teacher feedback is meaningful for learning and leads to deeper thinking and more comprehension (Chin, 2006; Erdogan & Campbell, 2008; Voerman et al., 2012). Teacher feedback facilitates both *comprehension cues* and *course demands cues*. Due to teacher feedback, students improve their metacognitive accuracy (Callender, Franco-Watkins & Roberts, 2016). The studies of Bachman and Bachman (2011), Bartsch and Murphy (2011), Mayer et al. (2009), Renner and Renner (2001) and Yourstone et al. (2008) all showed that teacher feedback during formative assessments with SRS has a positive impact on comprehension of course material and leads to higher scores on summative assessments. Moreover, students who are too inhibited to ask for help, or students who do not realize they need help, receive valuable information from their teacher without having to ask for it (King & Joshi, 2008). They receive feedback on how and when to use problem solving strategies and learn skills needed to answer the question. In line with this, the studies of Buil et al. (2019), Hattie and Timperley (2007), Nicol and Boyle (2003) and Shute (2008) all stated

that immediate teacher feedback after answering a question eliminates incorrect ways of thinking, and helps students to be aware of what they should focus their learning on, which motivates them to reinforce their learning. Here, teacher feedback may provide students with *motivation to learn cues*. According to Hattie and Timperley (2007), Koka and Hein (2006) and Shute (2008) teacher feedback has a powerful impact on students' motivation and willingness to continue their efforts to improve. Teacher feedback may trigger students to study or review one aspect of the material more than the other. This studying is determined internally by interest or externally by the emphasis the teacher place on certain aspects. Depending on the value they attach to the final grade on a test, their motivation determines which part of the course material they focus on (Efklides et al., 2018). However, there are also differences between groups of students. The study in Chapter 4 in this dissertation showed that teacher feedback during formative assessments with SRS motivates girls in particular to learn. Here, we argued that girls have a positive attitude to teacher feedback and are aware about the quality of this feedback, which improves their perceptions of competence and metacognitive skills.

#### 5.6.4 Peer discussions

In a more extended method of formative assessments with SRS, teacher feedback can be complemented with peer discussions.<sup>21</sup> Here, peer discussions serve as a fourth prompt for students (Figure 5.2). During peer discussions, students interact with their peers, and explain and justify their thinking to arrive at an answer (Mazur, 1997). Peer discussions have the aim of facilitating students with *comprehension cues* (Cortright et al., 2005; Mazur, 1997; Knight & Wood, 2005; Smith et al., 2009). These discussions stimulate the students providing the explanations and the recipient students to integrate and combine new knowledge and skills with existing knowledge and skills (Chi et al., 1994; Perez et al., 2010). Students usually explain

---

<sup>21</sup> Peer discussions can take place in several ways. For example, one way is that students first answer a multiple-choice question individually, discuss this question with their peers and submit the new (and potentially changed) answer again individually. After voting, the teacher shows a histogram and explains the question (e.g. Mazur, 1997; Molin, Haelermans, Cabus & Groot, 2020). Another way is that students, after answering the multiple-choice question individually, see the histogram (but not the correct answer), after which they are asked to work in groups to reach consensus and to vote a second time individually (e.g. Perez et al., 2010). It is also possible that students do not answer the question individually, but in small groups. Here, students share their SRS, reach consensus, and answer the question jointly (e.g. McDonough & Foote, 2015).

their thoughts in a more accessible language than teachers do. Recipient students therefore prefer hearing explanations from their peers (who have a similar background and a similar level of cognitive skills) rather than from a teacher (Falchikov & Goldfinch, 2000; Topping, Smith, Swanson & Elliot, 2000). Students may be better at addressing misunderstandings, proposing relevant examples or offering alternative solution strategies than a teacher (Caldwell, 2007; Hoekstra & Mollborn, 2012; Nicol & Boyle, 2003). Additionally, allowing students extra time to discuss their answers with each other encourages them to think critically, weigh options better, and consider other solution strategies (Hoekstra & Mollborn, 2012; Smith et al., 2009). Prior studies showed that the ability to solve problems and ultimately to correctly answer a multiple-choice question increases after peer discussion (e.g. Crouch & Mazur, 2001; Knight & Wood, 2005; Mazur, 1997; Molin, Haelermans, Cabus & Groot, 2021; Smith et al., 2009), even in cases where both students were initially wrong (e.g. Molin et al., 2021; Singh, 2005; Smith et al., 2009). Chi et al. (1994) and Brooks and Koretsky (2011) showed that even students who originally answer questions correctly develop richer explanations as a result of peer discussions, leading to more comprehension. Tullis and Goldstone (2020) showed that discussions provide a more thorough testing of answers and ideas than answering questions on its own. They argued that peer discussion facilitates metacognitive processes of detecting errors and assessing the coherence of an answer. Peer discussions stimulate students to have control over their own thinking, having certain thoughts about themselves and their peers. In discussions, students will need to take charge and pay attention to their actions so that they create a desired impression on their peers (Efklides, 2011). In turn, peers ask questions about why a problem should be thought about or act upon in that way to solve the problem, prompting fellow students to think further about their own actions. Peer discussions may also facilitate *motivation to learn cues*, because these discussions allow students to practice explaining concepts to one another, discovering answers or solutions that they may not find alone, which motivates them to study further and broader (Yu et al., 2014; Zhonggen, 2017). The discussions create a learning environment in which students become actively engaged by and reflecting their understanding on their own cognition. Using a variety of problem solving strategies through interaction with peers promote metacognitive awareness as they are stimulated to use, coordinate, and monitor various skills (Özcan, 2016; Paris & Paris, 2001). Peer discussions in small groups may also provide *anxiety cues*, as it could help decrease students' anxiety (Eddy, Brownell,

Thummaphan, Lan & Wenderoth, 2015; Hoekstra, 2008). The explaining and debating during small peer-group discussions ‘narrows the distance’ between students and make them feel at ease (Trees & Jackson, 2007; Yu et al., 2014), while they give students the opportunity to take a ‘laid back attitude’, allowing a laugh and a joke (Hoekstra, 2008). Ulbig and Notman (2012) showed that when students are asked to discuss their answers, shy or quiet students are more willing to explain how they arrived at an answer or are more willing to discuss other answer options than more confident students.

### 5.6.5 Isomorphic questions

During SRS use, teachers usually ask one multiple-choice question per concept, while students are expected to be able to apply a concept in different contexts (Reay, Li & Bao, 2008). Answering only one multiple-choice question per concept may often not be sufficient to alert students to the variety of problems that can arise when concepts are applied (Porter et al., 2011) or to indicate if they really understood the concept. A possible solution for this is to provide students with similar (isomorphic) questions, so that they can practice more per concept and assess their mastery in this concept (Porter et al., 2011; Smith et al., 2009). Isomorphic questions are two different questions that assess the same understanding of concepts, but with different numerical values in different contexts (Smith et al., 2009). Isomorphic questions may be a fifth prompt for students (Figure 5.2) which provides them with *comprehension cues*, as they enable the students to verify if they do indeed comprehend course content due to teacher feedback or peer discussions (Zingaro & Porter, 2014). Comprehension is demonstrated when students first individually answer a question incorrectly, and then after teacher feedback or peer discussions, answer a follow-up isomorphic question correctly (Barth-Cohen et al., 2016; Molin et al., 2021; Porter et al., 2011; Smith et al., 2009; Zingaro & Porter, 2014). Reinforced understanding also occurs when the isomorphic question is answered in a less time than the concept question posed earlier. Comprehension is lost when students answer a first question correctly and a follow-up isomorphic question incorrectly. Porter et al. (2011) state that in this latter case, students lack a general model of the underlying concept. Whether or not a follow-up isomorphic question is answered correctly is important for students to assess whether they comprehend the concept of the course content (Porter et al., 2011; Smith et al., 2009; Zingaro & Porter, 2014). Gick and Holyoak (1983) suggest that some students can see parallels between isomorphic

questions with very explicit hints. These students acknowledge that comparisons need to be made between questions, making it likely that they are engaged in meaningful metacognitive reflection.

## 5.7 Utilization problems

Note that it is not self-evident that all the five aforementioned prompts lead to the use of more diagnostic cues. By offering prompts at 'wrong' moments or in a 'wrong' order, cues can be insufficiently recognized by students and therefore not predictive for subsequent test performance (Anderson & Thiede, 2008). For example, peer discussion prompts do not always lead to diagnostic comprehension cues if they are conducted in a 'wrong' order. This is not so much due to the peer discussion activity itself, but rather to the precise moment in time that the peer discussion takes place. Showing voting results (such as offering histogram feedback prompts) before peer discussion prompts bias in students' discussions and diminishes the diagnosticity of comprehension cues. Perez et al. (2010) showed that students who discuss their votes with their peers after seeing a histogram are more likely to switch from a less common to the most common response than students who discuss their votes without seeing the histogram. Because of an unconscious desire of students to conform, students will choose the visibly most common answer rather than constructing a correct answer through discussions with their peers. Perez et al. (2010) stated that a histogram should be shown only after the peer discussion. It is also possible that peer discussion prompts lead to more incorrect reasonings, or to discussions of ideas that do not relate to the answer options of the multiple-choice questions (James & Willoughby, 2011; Knight et al., 2017). As a result, students get a wrong or incomplete understanding of the problem, lowering the diagnosticity of the cues produced. To ensure that students receive correct feedback that indeed improves their accuracy of monitoring judgments with diagnostic cues, one possibility is to complement peer discussion prompts with teacher feedback prompts (Nielsen, Hansen & Stav, 2013; Smith et al., 2011; Zingaro & Porter, 2014).

## 5.8 Conclusion

We have developed a conceptual framework to identify relationships between factors that might be responsible for influencing metacognition when students are formatively assessed with SRS. The monitor and control model of Nelson and Narens (1990) and the cue-utilization framework of Koriat (1997) were the starting points to develop a conceptual framework that uses prompts in formative assessments with SRS to enhance the utilization of diagnostic cues. Previous literature demonstrated that it is very difficult for students to judge their own knowledge correctly. The conceptualization of the relationships between prompts and diagnostic cues was presented in order to gain further insight how students can improve the accuracy of monitoring judgments and their metacognitive skills. We discussed the fact that a formative assessment may contain multiple prompts, with the variety and number of prompts depending on how the assessment is organized. The individual prompts may be helpful for students, as they aid them to identify one or more diagnostic cues that are predictive of subsequent learning and performance; one more than the other. When used jointly, these prompts might be even more effective. The framework assumes that more prompts during an assessment could lead to more diagnostic cues, an improvement of accuracy of monitoring judgments and an enhancement of metacognitive skills. These prompts and diagnostic cues provide teachers with directions as to how they can organize a formative assessment when using SRS, and they highlight the effects of their choices upon the types of behaviors and thought processes of their students.

We should note that it is not self-evident that all aforementioned prompts lead to use of more diagnostic cues, as not every implementation of an active and collaborative activity will lead to the generation and use of diagnostic cues. Formative assessment with SRS will not always improve metacognitive accuracy, as this depends on how a poll is conducted.

The developed framework is focused on formative assessment with SRS, but the monitor and control model of Nelson and Narens (1990) and the cue-utilization framework of Koriat (1997) are valid to all formative assessments i.e. also formative assessments without SRS. Therefore, most elements of the framework are relevant to all general formative assessments. However, the added value of formative testing with SRS is that students can vote anonymously and receive histogram feedback on top of teacher feedback and peer discussion. It is the



combination of these prompts and the use of technology that provides a sense of control over the learning process that makes formative assessment with SRS valuable for influencing metacognition. Unfortunately, metacognition is a recent area of research in formative assessment with SRS that is not yet well understood. This means that limited empirical research has been conducted in this area. For this reason, it is unclear (1) which prompts from polling influence which cues that finally lead to changes in metacognitive control decisions, and (2) how prompts influence metacognitive accuracy. More empirical research is needed to make predictions or recommendations regarding how formative assessment with SRS can be implemented to improve metacognitive monitoring accuracy.





## Chapter 6

### Conclusion and Discussion

## 6.1 Conclusion and discussion

Formative assessments are effective interventions in science education for improving academic performance. It is a method that teachers use during their lessons that provides feedback to modify ongoing teaching and learning to improve academic performance. As providing feedback has a large impact on student performances (Hattie, 2009), it is important to understand how formative assessments influence learning in everyday science classrooms. The aim of this dissertation was to provide evidence on how the way feedback is provided affects anxiety, motivation, performance and metacognitive awareness of learning in science courses. This led to the general research question formulated in the introduction of this dissertation: *'What types of digital formative assessments improve student learning?'* This general question resulted in a number of sub-questions that were examined in quantitative studies in *Chapters 2 through 4* and in a conceptual study in *Chapter 5* in this dissertation.

*Part II* of this dissertation focused on student performance in a final exam and on learning gains during formative assessments. While there are already numerous studies on the impact on student outcomes of formative assessments, experimental research on reasons for observing these outcomes is lacking.

*Chapter 2* evaluated the effects of repeated formative assessments with student response systems (SRS) compared to traditional teaching. Over a period of 17 weeks, a randomized experiment was carried out among 139 physics students in one school in Dutch secondary education. The results show that repeated formative assessments with SRS lowers physics anxiety and increases academic performance in a final exam.

*Chapter 3* examined whether feedback strategies (teacher feedback and peer discussions combined with teacher feedback) during formative assessments with SRS affect comprehension when answering isomorphic multiple-choice questions. A randomized experiment was conducted among 527 physics students from six Dutch secondary schools. The intervention shows that feedback strategies affect learning gains in comparison with students who receive no feedback.

*Part III* of this dissertation sheds light on the issue to what extent and how formative assessments with SRS affect metacognitive awareness. The role of metacognition in improving academic performance is increasingly noted, but insights into how and to what extent metacognition is influenced by formative assessments are lacking.

*Chapter 4* examined the effect of feedback strategies (teacher feedback and peer discussions combined with teacher feedback) on metacognitive awareness and motivation when students use SRS during formative assessments. A randomized experiment was carried out over a period of 10 weeks among 633 physics students from six Dutch secondary schools. The results indicate that the way feedback is provided affects both metacognitive awareness and motivation.

*Chapter 5* presented a conceptual framework of monitoring and control in formative assessments with SRS, highlighting factors that may influence metacognitive awareness. The framework assumes that prompts (or activities) during a formative assessment led to diagnostic cues, an improvement of accuracy of monitoring judgments, and an enhancement of metacognitive awareness.

The findings and conclusions from previous chapters are combined in this final chapter. To provide a brief overview, Table 6.1 shows all research questions posed and corresponding results in this dissertation. The gained insights are compiled into five key conclusions. All key conclusions are printed in italics. We combine the findings by discussing the key conclusions, the limitations of each study and the practical implications. We conclude this chapter with recommendations.

Table 6.1: Main findings of all research questions

Research question	Main findings
- Do formative assessments reduce anxiety in physics compared to traditional teaching? (Chapter 2)	Formative assessments significantly reduce anxiety in physics compared to traditional teaching. It corresponds to a small to medium effect size.
- Do formative assessments improve academic performance in physics compared to traditional teaching? (Chapter 2)	Formative assessments significantly improve academic performance in physics. It corresponds to a medium effect size.
- Does anxiety work as a mediating factor for the effect of formative assessments on academic performance? (Chapter 2)	Anxiety in physics is a mediator. Formative assessments reduce anxiety and in turn significantly affect academic performance.
- Do technology/polling system supported assessment activities enhance learning gains? (Chapter 3)	Teacher feedback, whether or not combined with peer discussions positively affects learning during a formative assessment compared to students who receive no feedback.
- Are learning gains modified by peer discussions? (Chapter 3)	Students can learn from peer discussions. Students who correct an initially incorrect concept question after peer discussions, answer proportionally more isomorphic questions correctly than students in the control condition who understood the concept and answer a concept question correctly.
- What is the effect of teacher feedback whether or not combined with peer discussions on metacognition? (Chapter 4)	Peer discussions combined with teacher feedback significantly positively affect metacognition. There is no significant effect for teacher feedback on metacognition.
- What is the effect of teacher feedback whether or not combined with peer discussions on motivation? (Chapter 4)	Peer discussions combined with teacher feedback significantly positively affect motivation. There is no significant effect for teacher feedback on motivation.
- Are there differential effects among girls and boys? (Chapter 4)	Girls increase their metacognitive skills and motivation via both teacher feedback and peer discussions combined with teacher feedback, while boys increase their metacognitive skills and motivation only via peer discussions combined with teacher feedback.
- Are there differential effects among students with high-, middle-, or low- metacognitive skills? (Chapter 4)	Students with low metacognitive skills benefit significantly more from peer discussions on top of teacher feedback than students with high metacognitive skills.
- Does the effect of feedback on metacognition run through its influences on motivation? (Chapter 4)	The effect of peer discussions combined with teacher feedback on metacognition runs partly through its influences on motivation.
- Do prompts in formative assessments influence metacognition? (Chapter 5)	Prompts lead to diagnostic cues, which improve students' accuracy of monitoring judgments and an enhanced of metacognition.

1. *Providing feedback in formative assessments increases both short-term and long-term course understanding.*

The first conclusion deals with the effectiveness of providing feedback (1) in the short term in a formative assessment, and (2) in the long term in a final exam. Previous studies have demonstrated that feedback has positive effects on course understanding. Although there is no generally adopted model for how feedback increases course understanding (Van der Kleij, Feskens & Eggen, 2015), prior research has shown that students modify or enhance their understanding after receiving feedback.

In *Chapter 3*, we evaluated the short-term effect of feedback on course understanding during formative assessments with SRS, in which students answered pairs of multiple-choice questions. Each first concept question was followed by a second isomorphic (similar) question, where students received only teacher feedback in an individual approach, and peer discussions followed by teacher feedback in a cooperative approach. The results in this chapter show that receiving teacher feedback on concept questions, whether or not combined with peer discussions, positively affects course understanding in the short term compared to students who receive no feedback or do not discuss with their peers. The results imply effect sizes of  $d = 0.83$  for teacher feedback and  $d = 1.13$  for peer discussion combined with teacher feedback, and show that students who receive feedback apply their acquired knowledge of the course content to new isomorphic questions in the short term compared to students who do not receive feedback. Both effect sizes are large (Cohen, 1988) and according to Hattie and Timperley (2007) are equivalent to the effect size of receiving feedback from teachers, students and peers on how to solve problems more effectively ( $d = 0.95$ ). The studies of Smith et al. (2009) and Egelandstad and Krumsvik (2017) showed similar results for teacher feedback and peer discussions on top of teacher feedback.

In *Chapter 2*, we examined the effects of repeated formative assessments with SRS on academic performance in a final exam, which allow us to identify the effects of feedback in the long-term. Here, treated students were formatively assessed with SRS at the end of each physics class, while students in the untreated condition did not receive these assessments. The students in the treated condition were allowed to discuss their responses, but teachers indicated that most treated students answered the questions individually and in silence. The results show that formative assessments improve academic performance in physics in an



exam compared to traditional teaching. This means that students who are daily formatively assessed and receive feedback on their answers gain more course understanding in the long-term than students who do not receive these assessments and feedback on questions. This long-term course understanding corresponds to an effect size of  $d = 0.34$ .<sup>22</sup> This is in line with the meta-analysis of Hattie (2009) and the studies of Black and Wiliam (1998a & 1998b) who both showed that formative assessments have positive effects on academic performance. The studies of Bachman and Bachman (2011) and Lin et al. (2011) found similar results that students who receive feedback with SRS outperform students in a control group who receive no feedback. Our result is comparable with average gains in effect size in active learning strategies; about 0.31 in physics (Freeman et al., 2014). With formative assessments, students become familiar with an exam preparation activity in which they answer questions that require problem-solving skills similar to those in a future summative assessment. Simultaneously, the weekly assessments divide the course content into small, manageable units, giving students more timely immediate and specific feedback about why an answer is correct or incorrect.

The combined results of *Chapters 2* and *3* show that providing feedback in formative assessments increases both short-term and long-term course understanding. We should note, however, that some caution is required, as the effect of improved long-term course understanding may have been enhanced by the repeated assessments; the so-called ‘testing effect’ (Karpicke & Blunt, 2011; McDaniel, Roediger & McDermott, 2007). The study of Roediger and Karpicke (2006) showed that long-term course understanding is enhanced when one or more assessments are included during learning. The studies of Campbell and Mayer (2009) and Mayer et al. (2009) demonstrated that formative assessments improve students’ exam performance, and the authors linked the results to repeated formative testing.

---

<sup>22</sup> The long-term effect (Chapter 2) is smaller than the short-term effect (Chapter 3). The reduced long-term effect can be explained, for example, by the long time span between the formative assessments and the final exam, which reduces, for example, students' ability to recall problem solving skills from memory. However, we should also note that the effect sizes are difficult to compare with each other. A main reason is that the control conditions in both studies are not similar to each other. For example, the control condition in the study of Chapter 2 did not receive formative assessments but only traditional teaching, while the control condition in the study of Chapter 3 were formatively assessed but received no feedback. Another reason is that the conditions under which the effects were measured (summative of nature versus formative of nature) are different, so other elements (e.g. anxiety) may affect the results. For these reasons, the effect sizes are difficult to compare.

For these reasons, in addition to feedback, the ‘testing effect’ could also be, in part, a possible explanation for the long-term course understanding.

*2. Students learn from peer discussions and thereby improve their metacognitive awareness.*

The second conclusion builds on the results of *Chapter 3*. In this chapter, we plotted flowcharts of students’ response patterns of paired multiple-choice questions in the treated and control conditions. The findings suggest that students in the cooperative condition who could learn from peer discussions, because they initially answered their first concept question incorrect, but corrected it after peer discussion, learn from peer discussions. These students answer proportionally more isomorphic questions correctly than students who receive no feedback or who do not discuss with their peers, but who answered the concept question correct at the beginning. We assume that students indeed learn from peer discussions and do not simply copy the same answer of their more skilled or more dominant peer. These findings are similar to the studies of Smith et al. (2009) and Egelandstad and Krumsvik (2017) who both showed that students who do not understand the concept of a question can learn from their peers; students share and (re)construct information while building on their own concepts and on the ideas of their peers. This leads to an additional reinforcement of course understanding and a transfer of acquired knowledge to new follow-up isomorphic questions.

After showing that students can actually learn from peer discussions, we theorized in *Chapter 5* that peer discussions facilitate students with comprehension cues, which can improve students’ accuracy of monitoring judgments and an enhancement of metacognition. We noted that peer discussions allow students to practice explaining concepts to one another, and explore answers or solutions they might not find on their own, which encourage them to study further and more thoroughly. With the evidence of *Chapter 3* in mind, in *Chapter 4* we examined the effects of feedback strategies (teacher feedback or a combination of peer discussions and teacher feedback) on metacognition. We demonstrate that peer discussion combined with teacher feedback has a positive effect on metacognition compared to students who receive no feedback. This finding has an effect size of  $d = 0.40$  and is in line with the studies of Blasco-Arcas et al. (2013), Brady et al. (2013), Levesque (2011) and Mazur (1997). They stated that peer discussions stimulate in-depth understanding and develop reasoning and metacognitive skills that improve transfer of knowledge from one problem to another

problem. On the basis of our findings, we conclude that students learn from peer discussions and that they improve their metacognitive awareness.

*3. The influence of feedback on metacognition differs by gender and level of metacognitive skills.*

The previous conclusion stated that peer discussions combined with teacher feedback increase metacognitive skills. Our third conclusion is about the differential effects of subgroups of students on metacognition and is based on *Chapter 4*. The studies of Brady et al. (2013) and Jones et al. (2012) both suggested that there is a relationship between (1) gender and growth of metacognitive skills due to feedback strategies on the one hand, and (2) prior level of metacognitive skills and growth in metacognitive skills due to feedback strategies on the other hand. The first finding shows that there are different effects among girls and boys. Compared to boys, metacognitive skills increase for girls via both teacher feedback or a combination of peer discussions and teacher feedback, while for boys metacognitive skills increase only via peer discussions combined with teacher feedback. This result is in line with the studies of King and Joshi (2008) and Lorenzo et al. (2006) who both demonstrated that girls and boys benefit from feedback strategies, but that they approach learning differently. Girls are more aware of the quality of the teacher feedback and use it to monitor their own learning, while boys need to actively participate in peer discussions and explain their own thoughts to become more metacognitive aware.

The findings in *Chapter 4* also show that there are differential effects among subgroups of students with different metacognitive skills. Students with low metacognitive skills benefit significantly more from peer discussions on top of teacher feedback than students with high metacognitive skills. We assume that this is because low-metacognitive students benefit from interactions with their high-metacognitive peers. This is confirmed by studies of Schraw et al. (2012) and Shapiro et al. (2017) who both showed that low-metacognitive students benefit from instruction and collaborating with a more experienced, metacognitive student. An additional explanation is that low-metacognitive students may benefit more from the 'application' and 'analyzing' level than high-metacognitive students. Based on these results, we conclude that the influence of feedback on metacognition differs for both boys and girls, and low and high-metacognitive students.

4. *More short-term course understanding could lead to more metacognitive awareness when students discuss with their peers and receive feedback from their teachers.*

The fourth conclusion addresses the relationship between course understanding and metacognitive awareness. *Chapters 3* and *4* examined whether teacher feedback, or peer discussions combined with teacher feedback affect students' short-term course understanding (*Chapters 3*) and metacognitive awareness (*Chapters 4*). In *Chapter 3*, we demonstrated that both teacher feedback and peer discussions on top of teacher feedback affect course understanding in the short term compared to students who receive no feedback or do not discuss with their peers, while in *Chapter 4*, we showed that only peer discussions on top of teacher feedback increase metacognitive awareness. These combined findings point out that short-term course understanding leads to increased metacognitive skills only when students engage in peer discussions on top of teacher feedback. We conclude that course understanding after only passively listening to the teachers' feedback is sufficient to understand course content in the short term, but is insufficient to monitor and regulate learning in a way that it increases metacognitive skills. It is in fact the combination of teacher feedback with peer discussions that reinforces conceptual understanding and increase metacognitive awareness, as peer discussions create an environment in which students demonstrate, explain, discuss, and control their own thought processes. Moreover, students judge their peers' explanation and compare them to their own reasonings and answers (Kollar & Fischer, 2010; Topping, 1998). In explaining their own thoughts and discussing and checking the feedback from their peers, students improve their critical thinking skills and develop metacognitive awareness (Lynch, McNamara & Seery, 2012; Smith, Cooper & Lancaster, 2002; Tsai, Lin & Yuan, 2002).

5. *Student feelings of anxiety in physics and motivation to learn mediate the effects of formative assessments.*

The fifth conclusion focuses on the effects of anxiety in physics and motivation to learn during formative assessments. In *Chapter 2*, we examined the effects of repeated formative assessments with SRS on student anxiety in physics and academic performance compared to traditional teaching. The results show that repeated formative testing has a significant positive effect on both anxiety in physics and academic performance compared to traditional teaching.

These results are confirmed by the studies of Brady et al. (2013) and Yu et al. (2014) who found that anonymity in classrooms by SRS releases students' anxiety. McDaniel et al. (2011) showed that repeated formative assessments reduce anxiety. The ability to answer questions anonymously, and the certainty that students will not be judged by their responses in the form of a grade, provides active participation and involvement, and releases anxiety. However, improved performances as a consequence of repeated formative testing do not stand on their own, as a mediation analysis shows that anxiety in physics plays a significant role; anxiety completely mediates the effect of formative assessments on academic performance. Thus, formative assessments significantly reduces anxiety, which in turn also significantly affects academic performance. This mediating effect of anxiety has not been examined before in studies focusing on formative testing. A mediation effect is also studied in *Chapter 4*. This chapter sheds light on how feedback strategies, which consists of teacher feedback or a combination of peer discussions and teacher feedback, affect students' motivation and metacognition. Balsco-Arcas et al. (2013) and Brady et al. (2013) reported that peer discussions in formative assessments affect metacognitive awareness, while the studies of Sun and Hsieh (2018) and Camacho-Miñano and Del Campo (2016) showed that assessments with SRS scaffold the development of students' motivation. The findings in *Chapter 4* indicate a significant positive effect of peer discussions combined with teacher feedback on both metacognition and motivation. However, the improvement of metacognition is partly mediated through motivation when students receive teacher feedback in combination with peer discussions. A likely explanation for this finding is that peer discussions make formative assessments more interactive and interesting and make students more proactive and focused when answering multiple-choice questions, which leads to more motivation to learn and, indirectly, to more metacognitive skills. Based on the findings in *Chapter 2* and *Chapter 4*, we conclude that feelings of anxiety in physics and feelings of motivation to learn mediate the effects of prompts on academic performance and metacognition, respectively. This mediation means that prompts, as discussed in *Chapter 5* also trigger cues like anxiety in physics and motivation to learn, which in turn influence metacognition and academic performance. This is in line with Efklides (2008), who stated that through 'sheer subjective' feelings such as motivation to learn cues and anxiety cues, students are aware of their beliefs and ideas about cognition (metacognition).

## 6.2 Study limitations and future research

The experimental and conceptual studies in this dissertation have provided important insights and filled several gaps in our knowledge of formative assessments with SRS. Nevertheless, the studies in this dissertation have limitations, which are discussed below along with their implications for further research.

*Chapter 2* evaluated the effects of repeated formative assessments with SRS on anxiety and academic performance compared to traditional teaching. The main limitation of the study in *Chapter 2* is that it does not allow us to determine whether the effects of a better performance and less anxiety compared to traditional teaching are mainly due to the repeated testing or to the feedback provided, based on the answers of the multiple-choice questions. Taking assessments itself prompts students to retrieve information from their long-term memory, which means that students also benefit from the ‘testing effect’. Most studies conducted in the past were unable to establish a distinction between the effect of testing and receiving feedback (e.g. Campbell & Mayer, 2009; Mayer et al., 2009; Shapiro & Gordon, 2012). Further research should focus on separate aspects, or on a different share of time spent on each aspect, to determine which one of those is most important for the significant positive results. Further research that includes a larger sample should reveal how robust the effect sizes are to the design and context of the study.

*Chapter 3* examined whether feedback strategies during formative assessments with SRS affect comprehension and course understanding when answering isomorphic multiple-choice questions. With regard to this study, the main limitation is that we did not attempt to compare only peer discussion as a separate condition between a concept question and an isomorphic question. In our study, students in the cooperative condition received teacher feedback after peer discussions. If we had a separate peer discussion condition, we could compare this condition with a control condition and drawn more strong conclusions about the contribution of peer discussions to learning gains. Another limitation is that students in the cooperative condition were asked to discuss their responses on a concept question with a student next to them in class. In itself, this need not be a problem, but since students were free to choose a seat in class, they probably tend to sit next to friends or classmates who share the same interests or who have a similar level of understanding, resulting in situations where students

are more likely to agree in peer discussions than expected. For this reason, further research should examine the extent of learning gains when students are randomly paired. We expect that peer discussions will then lead to even more verbal explanations and new knowledge, so that the results will be even larger than what we find in the present study. A third limitation is that we used SRS that allow students to answer only multiple-choice questions with a limited number of possible answers. As students have to choose from a range of answers, SRS restrict teachers from posing questions that require students to evaluate issues or express their creative ideas. Multiple-choice questions cannot assess the highest levels (the evaluating and creating levels) in the Anderson and Bloom's taxonomy (Anderson & Bloom, 2001) as all answers are presented, preventing students from coming up with new models or hypotheses on their own (Crowe, Dirks & Wenderoth, 2008). Thus, with respect to higher-order thinking, teachers using SRS in formative assessments are limited to asking questions at the application and analysis level (as occurred in Chapter 3).

*Chapter 4* examined the effects of feedback strategies during formative assessments with SRS on motivation and metacognitive awareness. With respect to this study, an important limitation is that we did not study the actual quality and quantity of the peer-to-peer conversations. This is because students have different levels of metacognitive skills, motivation and knowledge that are influenced by time of day and course content. These variables can all impact the quality and quantity of peer-to-peer conversations. Further research should analyze conversations and determine the extent to which the content of conversations leads to more conceptual understanding of the course content. Research about this is important, because teachers can then decide how much time they should allow students to discuss with peers thoughtfully while still maximizing class time.

*Chapter 5* presented a conceptual framework, highlighting factors that may influence metacognitive awareness. While we described that our lists of prompts and diagnostic cues are likely to be important ones when using SRS, an important limitation is that there may also be other prompts and diagnostic cues that have not yet been mentioned in our framework. Our study could serve as a basis for further research. Furthermore, empirical research is needed to test our framework, in order to find out whether there are other prompts and cues that are predictive of test performance. As our focus was predominantly on the monitoring part of the framework, further research is also needed to investigate the link between

monitoring judgments and control. This should help students determine which judgments lead to control decisions and which control processes in turn lead to an actual improvement of test performances.

### 6.3 Practical implications

In everyday classroom practice, there is a need for teachers to provide assessments that help them to improve their instruction. Traditional strategies of formative assessments (e.g. checking assignments or discussing homework) are being used less and less, because they lack the ease of use and organization that SRS provide. By using SRS, teachers are able to gain insight into student understanding in a brief period of time and they can easily organize assessments in which students are actively encouraged to answer questions, participate in peer discussions, ask questions to the teacher and reflect on their own learning. At the same time, teachers build on student ideas, provide meaningful feedback to move students forward in their learning and make instructional decisions about subsequent lessons. Depending on the purpose, teachers may choose to formatively assess students at the start of the class (e.g. to retrieve knowledge or to check course content that needs to be repeated), halfway through the class (e.g. to keep students' attention), or at the end of class (e.g. to check whether the course content has been understood). When teachers use formative assessments with SRS, they use prompts. This dissertation shows that formative assessments contain multiple prompts, with the number and variety of prompts depending on how teachers organize assessments. In a limited time for testing, a teacher may choose to provide teacher feedback only. This dissertation shows that compared to traditional teaching, this form of repeated testing is effective for students to improve their physics performance and reduce their anxiety in physics. Here, however, it is unclear whether the improved performance is primarily a result of teacher feedback or the 'testing effect', since testing course material can have a positive effect on memory retention (Roediger & Butler, 2011). Nevertheless, the effects found are medium for improving grades in a final exam and small to medium for reducing anxiety in physics. From a practical perspective, *Chapter 2* in this dissertation argues that substantial effects can be achieved in upper secondary school physics education when teachers have limited time (10 to 15 minutes) for formative testing.



If teachers have more time, they may choose to have students engage in peer discussions. When peer discussions and teacher feedback are combined, students first answer a multiple-choice question individually, discuss this question with their peers and then re-vote the question with a new, potentially revised answer, after which the teacher displays the histogram and explains the correct answer. *Chapter 3* shows that peer discussions help students to choose a correct answer and assumes that students do indeed learn from peer discussions. Students reveal that peer discussions break the monotony of passive listening to the teacher and encourage them to explain their own reasoning and listen to what others have to say. Most improvements in answering questions correctly occur when peer discussions are combined with teacher feedback. With these findings, this study shows that peer discussions matter. Having peer discussions is not a waste of time but an added value to the process of formative testing. This is further shown in *Chapter 4*, which demonstrates that peer discussions on top of teacher feedback have a positive effect on metacognitive skills and motivation. More metacognitive awareness enables students to think about, understand, and monitor their learning; skills that are critical in learning sciences, such as physics (Mota et al., 2019; Taasobshirazi et al., 2015). Additional teacher feedback remains important and should be given after peer discussions, as it is possible that students discuss incorrect ideas, or answer questions that are inconsistent with the ideas they discuss. For this reason, teachers should use formative assessments with SRS in conjunction with peer discussions and teacher feedback to engage students in deeper metacognitive monitoring and developing metacognitive skills. In doing so, teachers accommodate both girls and boys, and low-metacognitive students. A requirement is that teachers take the group composition of students during peer discussions into account, so that low-metacognitive students are grouped with high-metacognitive students.

## 6.4 Recommendations

Taking assessments from students that are focused on the progress of learning, such as formative assessment, is potentially one of the most important tasks of teachers. In fact, enhancing knowledge about formative assessments increases the quality of learning. However, it is plausible that this will cause pressure to teachers. To increase the importance of knowledge about formative assessment we formulate a number of recommendations that

can improve the daily testing practice. This dissertation shows that peer discussions during formative assessment really matters. Peer discussions enhance the learning process. Yet, it is not obvious to all teachers to allow students to discuss with their peers during a formative assessment, because they believe that students do not have enough knowledge to judge the correctness of a reasoning of their peer. For these teachers, peer discussion is a waste of time (Nicol & Boyle, 2003). However, students do indeed learn from peer discussions. We recommend that teachers make peer discussions a standard part of formative assessments. To ensure that students engage in critical discussions, a teacher can walk through the class meanwhile, listen to students and ask them to clarify reasons behind their answers. The 'lost' time is finally compensated because the teacher has less time to spend giving feedback.

A second recommendation is to use isomorphic questions. Most (quasi-) experimental studies show that teachers usually ask one multiple-choice question per concept, while students are expected to be able to apply a concept in different contexts. Answering only one multiple-choice question per concept may often not be sufficient to alert students to the variety of problems that can occur when concepts are applied. These paired isomorphic questions allow students to assess whether they have actually understood the course material. Understanding occurs when students answer a subsequent isomorphic question correctly after peer discussions and teacher feedback. This means that teachers need to have a database with a sufficient number of isomorphic questions so that they can select questions in an upcoming formative assessment if it is found that students did not yet fully understand the course material. Compiling questions for a database is a time-consuming task that requires knowledge and experience. The questions should, in fact, require several thinking steps and should assess a sufficiently deep conceptual understanding.

A third recommendation is that teachers collaborate in this process, as assessment gains in quality when teachers work together (Van der Klink, 2012). Collaboration should not only take place in writing and composing questions, but also in assessing the level and depth of questions. Creating sufficient professional space is a crucial part of this. Training teachers in test-making is often neglected in schools, but when school leaders facilitate teachers with time and training, this also improves teachers' proficiency in test-making.



# Chapter 7

## Impact Statement

## 7.1 Impact statement

Formative assessments are effective interventions in science education for improving academic performance. These assessments are interventions that prompt students to think about problems or course material, and aims to provide information to the teacher and student about the extent to which the material has or has not been mastered. In recent decades, formative assessments have gained increased attention because research shows that they can improve student learning. But too often the way formative assessments should occur is not consistent with what actually takes place in daily teaching practice, since teachers do a lot based on their gut feeling. So is the situation with formative assessments in science courses. A large number of science teachers do not realize that the way an assessment is organized and the way feedback is given affect students' feelings and learning. The aim of this dissertation is to provide evidence on how the way feedback is provided affects anxiety, motivation to learn, metacognitive awareness, and short-term and long term-course understanding in physics courses.

This dissertation shows that formative assessments contain multiple prompts, with the number and variety of prompts depending on how teachers organize assessments. In a limited time for testing, a teacher may choose to provide teacher feedback only. Based on an experimental study in everyday physics classrooms, this dissertation demonstrates that repeated formative assessments reduce anxiety in physics and improve academic performance in physics compared to traditional teaching. This implies that repeated formative assessments can make students feel more at ease, which contributes to more academic performance. For this reason, we argue that substantial effects can be achieved in upper secondary school physics education when teachers have limited time (10 to 15 minutes per class) for formative testing.

If teachers have more time, they may choose to have students engage in peer discussions. The results in this dissertation show that teacher feedback, whether or not combined with peer discussions, positively affect learning gains in comparison with students who do not receive any kind of feedback. The largest learning gains occur when peer discussions are followed by teacher feedback. Students learn from peer discussions and do not simply copy the same answer of their more skilled or more dominant peer. We argue that students can

indeed learn from peer discussions. With this evidence in mind, this dissertation shows that students who receive teacher feedback combined with peer discussions increase their motivation and metacognitive awareness in comparison to students who do not receive any kind of feedback. More metacognitive awareness enables students to think about, understand, and monitor their learning; skills that are critical in learning sciences, such as physics. However, girls differ from boys, as girls increase their metacognitive skills and get more motivated through teacher feedback (whether or not in combination with peer discussions), while boys increase their metacognitive skills and motivation only through peer discussions combined with teacher feedback. Differential effects are also observed for students with different metacognitive skills, as low-metacognitive students benefit more from peer discussions on top of teacher feedback than high-metacognitive students. We conclude that teacher feedback alone might be insufficient for students to improve learning, but that additional teacher feedback remains important and should be given after peer discussions, as it is possible that students discuss incorrect ideas, or answer questions that are inconsistent with the ideas they discuss. For this reason, teachers should use formative assessments with SRS in conjunction with peer discussions and teacher feedback to engage students in deeper metacognitive monitoring and developing metacognitive skills.

With the results obtained, this dissertation provides knowledge to the existing literature on the effects of feedback strategies in formative assessments with SRS on students' science anxiety, motivation, metacognitive awareness and performance. It also contributes to the literature with the development of a conceptual framework that identifies relationships between factors that may be responsible for influencing metacognition. Teachers benefit from the studies in this dissertation, as they give them insight that formative assessments are more than just weekly assessments. These assessments increase students' motivation to learn and reduce students' anxiety; feelings that influence metacognition and academic performance. The studies provide teachers with evidence that easy to organize assessments have large learning improvements. Receiving feedback on concept questions, positively affects course understanding in the short term. The results imply large effect sizes for teacher feedback ( $d = 0.83$ ) and for peer discussion combined with teacher feedback ( $d = 1.13$ ), and show that students who receive feedback apply their acquired knowledge of the course content to similar questions compared to students who do not receive feedback. Both effect sizes are

equivalent to the effect size of receiving feedback from teachers, students and peers on how to solve problems more effectively ( $d = 0.95$ ; Hattie & Timperley, 2007). Students who are daily formatively assessed and who receive feedback on their answers gain also more course understanding in the long term than students who do not receive these assessments and feedback on their answers. This long-term course understanding corresponds to an effect size of  $d = 0.34$ .

We are aware that the studies in this dissertation are only relevant if teachers take note of the results and actually apply the findings in their classes. We have tried to keep our own studies close to everyday educational practice. In this way, we show that by brief formative testing every lesson or every week, the quality of teaching and learning can be increased. As stated previously, we recommend that teachers (1) make peer discussions a regular part of formative testing, and (2) use isomorphic questions that assess concepts in different contexts. The latter means that teachers should have a broad database of questions. This can be problematic, as creating enough questions takes a lot of time and expertise. It is an issue that schools themselves can take up, but it is also conceivable that this is a task for the Ministry or for publishers. However, cooperation between the various stakeholders and schools or teachers is important in this process, as the quality of formative assessments (and thereby teaching and learning) also depends on the quality of the questions.







## References

- Agarwal, P. K., D'Antonio, L., Roediger III, H. L., McDermott, K. B., & McDaniel, M. A. (2014). Classroom-based programs of retrieval practice reduce middle school and high school students' test anxiety. *Journal of Applied Research in Memory and Cognition*, 3(3), 131-139.
- Aleven, V. A., & Koedinger, K. R. (2002). An effective metacognitive strategy: Learning by doing and explaining with a computer-based cognitive tutor. *Cognitive science*, 26(2), 147-179.
- Anderson, N. J. (2002). The Role of Metacognition in Second Language Teaching and Learning. ERIC Digest.
- Anderson, L. W., & Bloom, B. S. (2001). A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives. Longman,.
- Anderson, M. C., & Thiede, K. W. (2008). Why do delayed summaries improve metacomprehension accuracy?. *Acta psychologica*, 128(1), 110-118.
- Andersson, C., & Palm, T. (2017). The impact of formative assessment on student achievement: A study of the effects of changes to classroom practice after a comprehensive professional development programme. *Learning and instruction*, 49, 92-102.
- Angrist, J. D., & Pischke, J. S. (2008). *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.
- Anthis, K. (2011). Is it the clicker, or is it the question? Untangling the effects of student response system use. *Teaching of Psychology*, 38(3), 189-193.
- Arkin, R. M., & Schumann, D. W. (1984). Effect of corrective testing: An extension. *Journal of Educational Psychology*, 76(5), 835, 843.
- Atkinson, R. K., Renkl, A., & Merrill, M. M. (2003). Transitioning from studying examples to solving problems: Effects of self-explanation prompts and fading worked-out steps. *Journal of educational psychology*, 95(4), 774.
- Avargil, S., Lavi, R., & Dori, Y. J. (2018). Students' metacognition and metacognitive strategies in science education. In *Cognition, Metacognition, and Culture in STEM Education* (pp. 33-64). Springer, Cham.
- Baars, M., & Wijnia, L. (2018). The relation between task-specific motivational profiles and training of self-regulated learning skills. *Learning and Individual Differences*, 64, 125-137.
- Bachman, L., & Bachman, C. (2011). A study of classroom response system clickers: Increasing student engagement and performance in a large undergraduate lecture class on architectural research. *Journal of Interactive Learning Research*, 22(1), 5-21.
- Balta, N., & Tzafilkou, K. (2019). Using Socrative software for instant formative feedback in physics courses. *Education and Information Technologies*, 24(1), 307-323.

## References

- Balta, N., Michinov, N., Balyimez, S., & Ayaz, M. F. (2017). A meta-analysis of the effect of Peer Instruction on learning gain: Identification of informational and cultural moderators. *International Journal of Educational Research*, 86, 66-77.
- Balta, N., Perera-Rodríguez, V. H., & Hervás-Gómez, C. (2018). Using Socrative as an online homework platform to increase students' exam scores. *Education and Information Technologies*, 23(2), 837-850.
- Bandura, A. (1993). Perceived self-efficacy in cognitive development and functioning. *Educational psychologist*, 28(2), 117-148.
- Bangert-Drowns, R. L., Kulik, C. L. C., Kulik, J. A., & Morgan, M. (1991). The instructional effect of feedback in test-like events. *Review of educational research*, 61(2), 213-238.
- Barnes, G. (1977). Scores on a Piaget-type questionnaire versus semester grades for lower-division college physics students. *American Journal of Physics*, 45(9), 841-847.
- Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51(6), 1173, 1182.
- Barth-Cohen, L. A., Smith, M. K., Capps, D. K., Lewin, J. D., Shemwell, J. T., & Stetzer, M. R. (2016). What are middle school students talking about during clicker questions? Characterizing small-condition conversations mediated by classroom response systems. *Journal of Science Education and Technology*, 25(1), 50-61.
- Bartsch, R. A., & Murphy, W. (2011). Examining the effects of an electronic classroom response system on student engagement and performance. *Journal of Educational Computing Research*, 44(1), 25-33.
- Batchelor, J. (2015). Effects of clicker use on calculus students' mathematics anxiety. *PRIMUS*, 25(5), 453-472.
- Beatty, I. D., Gerace, W. J., Leonard, W. J., & Dufresne, R. J. (2006). Designing effective questions for classroom response system teaching. *American Journal of Physics*, 74(1), 31-39.
- Becker, S. A., Brown, M., Dahlstrom, E., Davis, A., DePaul, K., Diaz, V., & Pomerantz, J. (2018). NMC horizon report: 2018 higher education edition. *Louisville, CO: Educause*.
- Benjamin, A. S., Bjork, R. A., & Schwartz, B. L. (1998). The mismeasure of memory: when retrieval fluency is misleading as a metamnemonic index. *Journal of Experimental Psychology: General*, 127(1), 55.
- Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in education: principles, policy & practice*, 18(1), 5-25.
- Betz, N. E. (1978). Prevalence, distribution, and correlates of math anxiety in college students. *Journal of Counseling Psychology*, 25(5), 441, 448.
- Biggs, J. B. (1985). The role of metalearning in study processes. *British journal of educational psychology*, 55(3), 185-212.
- Birenbaum, M. (1996). Assessment 2000: Towards a pluralistic approach to assessment. In *Alternatives in assessment of achievements, learning processes and prior knowledge* (pp. 3-29). Springer, Dordrecht.

- Birenbaum, M., & Nasser, F. (1994). On the Relationship between Test Anxiety and Test Performance. *Measurement and Evaluation in Counseling and Development*, 27(1), 293-301.
- Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: Beliefs, techniques, and illusions. *Annual review of psychology*, 64, 417-444.
- Black, P., & Wiliam, D. (1998a). Assessment and classroom learning. *Assessment in Education: principles, policy & practice*, 5(1), 7-74.
- Black, P., & Wiliam, D. (1998b). Inside the black box: Raising standards through classroom assessment. *Phi delta kappan*, 92(1), 81-90.
- Blasco-Arcas, L., Buil, I., Hernández-Ortega, B., & Sese, F. J. (2013). Using clickers in class. The role of interactivity, active collaborative learning and engagement in learning performance. *Computers & Education*, 62, 102-110.
- Bloom, B.S. (Ed.), Engelhart, M.D., Furst, E.J., Hill, W.H., & Krathwohl, D.R. (1956). *Taxonomy of educational objectives: The classification of educational goals*. Handbook 1: Cognitive domain. New York: David McKay.
- Bloom, H. S., Richburg-Hayes, L., & Black, A. R. (2007). Using covariates to improve precision for studies that randomize schools to evaluate educational interventions. *Educational Evaluation and Policy Analysis*, 29(1), 30-59.
- Bojinova, E., & Oigara, J. (2013). Teaching and Learning with Clickers in Higher Education. *International Journal of Teaching and Learning in Higher Education*, 25(2), 154-165.
- Boud, D., & Molloy, E. (2013). Rethinking models of feedback for learning: the challenge of design. *Assessment & Evaluation in Higher Education*, 38(6), 698-712.
- Brady, M., Rosenthal, J. L., Forest, C. P., & Hocevar, D. (2020). Anonymous versus public student feedback systems: metacognition and achievement with graduate learners. *Educational Technology Research and Development*, 68(6), 2853-2872.
- Brady, M., Seli, H., & Rosenthal, J. (2013). Metacognition and the influence of polling systems: how do clickers compare with low technology systems. *Educational Technology Research and Development*, 61(6), 885-902.
- Brady, M., Seli, H., & Rosenthal, J. (2013b). "Clickers" and metacognition: A quasi-experimental comparative study about metacognitive self-regulation and use of electronic feedback devices. *Computers & Education*, 65, 56-63.
- Brooks, B. J., & Koretsky, M. D. (2011). The influence of group discussion on students' responses and confidence during peer instruction. *Journal of Chemical Education*, 88(11), 1477-1484.
- Brotman, J. S., & Moore, F. M. (2008). Girls and science: A review of four themes in the science education literature. *Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching*, 45(9), 971-1002.
- Brown, P. C., Roediger III, H. L., & McDaniel, M. A. (2014). *Make it stick: The science of successful learning*. Boston: Harvard University Press.
- Buil, I., Catalán, S., & Martínez, E. (2016). Do clickers enhance learning? A control-value theory approach. *Computers & Education*, 103, 170-182.

## References

- Buil, I., Catalán, S., & Martínez, E. (2019). The influence of flow on learning outcomes: An empirical study on the use of clickers. *British Journal of Educational Technology*, 50(1), 428-439.
- Burns, D. J. (2004). Anxiety at the time of the final exam: Relationships with expectations and performance. *Journal of Education for Business*, 80(2), 119.
- Butler, A. C., Karpicke, J. D., & Roediger III, H. L. (2007). The effect of type and timing of feedback on learning from multiple-choice tests. *Journal of Experimental Psychology: Applied*, 13(4), 273.
- Butler, D. L., & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of educational research*, 65(3), 245-281.
- Butler, A. C., & Woodward, N. R. (2018). Toward consilience in the use of task-level feedback to promote learning. In *Psychology of Learning and Motivation* (Vol. 69, pp. 1-38). Academic Press.
- Caldwell, J. E. (2007). Clickers in the large classroom: Current research and best-practice tips. *CBE—Life Sciences Education*, 6(1), 9-20.
- Callender, A. A., Franco-Watkins, A. M., & Roberts, A. S. (2016). Improving metacognition in the classroom through instruction, training, and feedback. *Metacognition and learning*, 11(2), 215-235.
- Camacho-Miñano, M. D. M., & del Campo, C. (2016). Useful interactive teaching tool for learning: clickers in higher education. *Interactive Learning Environments*, 24(4), 706-723.
- Campbell, J., & Mayer, R. E. (2009). Questioning as an instructional method: Does it affect learning from lectures?. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, 23(6), 747-759.
- Carnaghan, C., & Webb, A. (2007). Investigating the effects of group response systems on student satisfaction, learning, and engagement in accounting education. *Issues in Accounting Education*, 22(3), 391-409.
- Carvalho, C., Santos, J., Conboy, J., & Martins, D. (2014). Teachers' feedback: Exploring differences in students' perceptions. *Procedia-Social and Behavioral Sciences*, 159, 169-173.
- Cassady, J. C. (2004). The impact of cognitive test anxiety on text comprehension and recall in the absence of external evaluative pressure. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, 18(3), 311-325.
- Cassady, J. C., & Gridley, B. E. (2005). The effects of online formative and summative assessment on test anxiety and performance. *The Journal of Technology, Learning and Assessment*, 4(1).
- Cassady, J. C., Budenz-Anders, J., Pavlechko, G., & Mock, W. (2001). The Effects of Internet-Based Formative and Summative Assessment on Test Anxiety, *Perceptions of Threat, and Achievement*.

- Cauley, K. M., & McMillan, J. H. (2010). Formative assessment techniques to support student motivation and achievement. *The clearing house: A journal of educational strategies, issues and ideas*, 83(1), 1-6.
- Champagne, A. B., & Newell, S. T. (1992). Directions for research and development: Alternative methods of assessing scientific literacy. *Journal of Research in Science Teaching*, 29(8), 841-860.
- Chang, S. N., & Chiu, M. H. (2005). The development of authentic assessments to investigate ninth graders' scientific literacy: In the case of scientific cognition concerning the concepts of chemistry and physics. *International Journal of Science and Mathematics Education*, 3(1), 117-140.
- Chen, J. C., Whittinghill, D. C., & Kadlowec, J. A. (2010). Classes that click: Fast, rich feedback to enhance student learning and satisfaction. *Journal of Engineering Education*, 99(2), 159-168.
- Chetty, R., Friedman, J. N., & Rockoff, J. E. (2011). *The Long-Term Impacts of Teachers: Teacher Value-Added and Student Outcomes in Adulthood*. NBER Working Paper no. 17699. National Bureau of Economic Research.
- Chi, M. T., De Leeuw, N., Chiu, M. H., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive science*, 18(3), 439-477.
- Chien, Y. T., Chang, Y. H., & Chang, C. Y. (2016). Do we click in the right way? A meta-analytic review of clicker-integrated instruction. *Educational Research Review*, 17, 1-18.
- Chien, Y. T., Lee, Y. H., Li, T. Y., & Chang, C. Y. (2015). Examining the effects of displaying clicker voting results on high school students' voting behaviors, discussion processes, and learning outcomes. *Eurasia Journal of Mathematics, Science and Technology Education*, 11(5), 1089-1104.
- Chin, C. (2006). Classroom interaction in science: Teacher questioning and feedback to students' responses. *International journal of science education*, 28(11), 1315-1346.
- Coca, D. M., & Sliško, J. (2017). Software Socrative and smartphones as tools for implementation of basic processes of active physics learning in classroom: An initial feasibility study with prospective teachers. *European Journal of Physics Education*, 4(2), 17-24.
- Cohen, H. D., Hillman, D. F., & Agne, R. M. (1978). Cognitive level and college physics achievement. *American Journal of Physics*, 46(10), 1026-1029.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* 2nd edn.
- Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Routledge.
- Cohn, S. T., & Fraser, B. J. (2016). Effectiveness of student response systems in terms of learning environment, attitudes and achievement. *Learning Environments Research*, 19(2), 153-167.
- Coleman, E. B. (1998). Using explanatory knowledge during collaborative problem solving in science. *Journal of the Learning Sciences*, 7(3-4), 387-427.

## References

- Cooksey, R. W., Freebody, P., & Wyatt-Smith, C. (2007). Assessment as judgment-in-context: Analysing how teachers evaluate students' writing. *Educational Research and Evaluation*, 13(5), 401-434.
- Cortright, R. N., Collins, H. L., & DiCarlo, S. E. (2005). Peer instruction enhanced meaningful learning: ability to solve novel problems. *Advances in physiology education*, 29(2), 107-111.
- Coutinho, S. A. (2006). The relationship between the need for cognition, metacognition, and intellectual task performance. *Educational research and reviews*, 1(5), 162-164.
- Covington, M. V. (1985). Test anxiety: Causes and effects over time. *Advances in test anxiety research*, 4, 55-68.
- Crins, J. (2002). Vragenlijst studievoorwaarden. *KPC Onderwijs Innovatie Centrum, 's-Hertogenbosch, Nederland*, 3.100.11.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *psychometrika*, 16(3), 297-334.
- Crossgrove, K., & Curran, K. L. (2008). Using clickers in nonmajors-and majors-level biology courses: student opinion, learning, and long-term retention of course material. *CBE—Life Sciences Education*, 7(1), 146-154.
- Crouch, C. H., & Mazur, E. (2001). Peer instruction: Ten years of experience and results. *American journal of physics*, 69(9), 970-977.
- Crouch, C. H., Watkins, J., Fagen, A. P., & Mazur, E. (2007). Peer instruction: Engaging students one-on-one, all at once. *Research-based reform of university physics*, 1(1), 40-95.
- Crowe, A., Dirks, C., & Wenderoth, M. P. (2008). Biology in bloom: implementing Bloom's taxonomy to enhance student learning in biology. *CBE-Life Sciences Education*, 7(4), 368-381.
- Dabbagh, N., Bass, R., Bishop, M., Costelloe, S., Cummings, K., Freeman, B., ... & Wilson, S. J. (2019). Using technology to support postsecondary student learning: a practice guide for college and university administrators, advisors, and faculty.
- De Bruin, A. B., & van Merriënboer, J. J. (2017). Bridging cognitive load and self-regulated learning research: A complementary approach to contemporary issues in educational research. *Learning and Instruction*, 51, 1-9.
- De Bruin, A. B., Dunlosky, J., & Cavalcanti, R. B. (2017). Monitoring and regulation of learning in medical education: the need for predictive cues. *Medical education*, 51(6), 575-584.
- De Gagne, J. C. (2011). The impact of clickers in nursing education: A review of literature. *Nurse Education Today*, 31(8), e34-e40.
- De Sousa, B. F. P. (2018). Engaging students in the evaluation process using co-creation and technology enhanced learning (CC-TEL). *CC-TEL. Leeds, UK*.
- DeBourgh, G. A. (2008). Use of classroom “clickers” to promote acquisition of advanced reasoning skills. *Nurse Education in Practice*, 8(2), 76-87.
- Deci, E. L., Ryan, R. M., & Williams, G. C. (1996). Need satisfaction and the self-regulation of learning. *Learning and individual differences*, 8(3), 165-183.

- DePasque, S., & Tricomi, E. (2015). Effects of intrinsic motivation on feedback processing during learning. *NeuroImage*, 119, 175-186.
- Desoete, A. (2008). Multi-method assessment of metacognitive skills in elementary school children: How you test is what you get. *Metacognition and Learning*, DOI.10.1007/s11409-008-9026-0.
- Dolin, J., & Krogh, L. B. (2010). The relevance and consequences of PISA science in a Danish context. *International Journal of Science and Mathematics Education*, 8(3), 565-592.
- Dori, Y. J., Mevarech, Z. R., & Baker, D. R. (2018). Cognition, metacognition, and culture in STEM education. *Innovations in Science Education and Technology*, 24, 386.
- Doucet, M., Vrins, A., & Harvey, D. (2009). Effect of using an audience response system on learning environment, motivation and long-term retention, during case-discussions in a large group of undergraduate veterinary clinical pharmacology students. *Medical Teacher*, 31(12), e570-e579.
- Dudaitė, J., & Prakapas, R. (2017). The experience of teachers in the application of activInspire interactive evaluation system in classroom: a case of teachers in Lithuania. *Informatics in Education*, 16(2), 181.
- Dunlosky, J., & Thiede, K. W. (1998). What makes people study more? An evaluation of factors that affect self-paced study. *Acta psychologica*, 98(1), 37-56.
- Dunlosky, J., Rawson, K. A., & Hacker, D. J. (2002). Metacomprehension of science text: Investigating the levels-of-disruption hypothesis. In J. Otero, J. A. Leo'n, & A. C. Graesser (Eds.), *The psychology of science text comprehension* (pp. 255-279).
- Dunning, D., Johnson, K., Ehrlinger, J., & Kruger, J. (2003). Why people fail to recognize their own incompetence. *Current directions in psychological science*, 12(3), 83-87.
- Eccles, J. S. (2005). Subjective task value and the Eccles et al. model of achievement-related choices. *Handbook of competence and motivation*, 105-121.
- Eddy, S. L., Brownell, S. E., Thummaphan, P., Lan, M. C., & Wenderoth, M. P. (2015). Caution, student experience may vary: social identities impact a student's experience in peer discussions. *CBE—Life Sciences Education*, 14(4), ar45.
- Edens, K. M. (2008). The interaction of pedagogical approach, gender, self-regulation, and goal orientation using student response system technology. *Journal of Research on Technology in Education*, 41(2), 161-177.
- Efklides, A. (2001). Metacognitive experiences in problem solving. In *Trends and prospects in motivation research* (pp. 297-323). Springer, Dordrecht.
- Efklides, A. (2006). Metacognition and affect: What can metacognitive experiences tell us about the learning process? *Educational research review*, 1(1), 3-14.
- Efklides, A. (2008). Metacognition: Defining its facets and levels of functioning in relation to self-regulation and co-regulation. *European Psychologist*, 13(4), 277-287.
- Efklides, A. (2011). Interactions of metacognition with motivation and affect in self-regulated learning: The MASRL model. *Educational psychologist*, 46(1), 6-25.



## References

- Efklides, A., Schwartz, B. L., & Brown, V. (2018). Motivation and affect in self-regulated learning: *Does metacognition play a role*. Handbook of self-regulation of learning and performance, 64-82.
- Egelandsdal, K., & Krumsvik, R. J. (2017a). Clickers and formative feedback at university lectures. *Education and Information Technologies*, 22(1), 55-74.
- Egelandsdal, K., & Krumsvik, R. J. (2017b). Peer discussions and response technology: short interventions, considerable gains. *Nordic Journal of Digital Literacy*, 12(01-02), 19-30.
- Egelandsdal, K., & Krumsvik, R. J. (2019a). Clicker Interventions at University Lectures and the Feedback Gap. *Nordic Journal of Digital Literacy*, 14(01-02), 70-87.
- Egelandsdal, K., & Krumsvik, R. J. (2019b). Clicker Interventions: Promoting Student Activity and Feedback at University Lectures. Tatnall (red.), *Encyclopedia of Education and Information Technologies*, 1-15.
- Elliot, A. J., & McGregor, H. A. (1999). Test anxiety and the hierarchical model of approach and avoidance achievement motivation. *Journal of Personality and social Psychology*, 76(4), 628.
- Engelen, J. A., Camp, G., van de Pol, J., & de Bruin, A. B. (2018). Teachers' monitoring of students' text comprehension: can students' keywords and summaries improve teachers' judgment accuracy?. *Metacognition and learning*, 13(3), 287-307.
- Entwistle, N., & Entwistle, D. (2003). Preparing for examinations: The interplay of memorising and understanding, and the development of knowledge objects. *Higher Education Research & Development*, 22(1), 19-41.
- Erdogan, I., & Campbell, T. (2008). Teacher questioning and interaction patterns in classrooms facilitated with differing levels of constructivist teaching practices. *International Journal of Science Education*, 30(14), 1891-1914.
- Evans, C. (2013). Making sense of assessment feedback in higher education. *Review of educational research*, 83(1), 70-120.
- Everson, H. T., Smodlaka, I., & Tobias, S. (1994). Exploring the relationship of test anxiety and metacognition on reading test performance: A cognitive analysis. *Anxiety, Stress and Coping*, 7(1), 85-96.
- Faber, J. M., Luyten, H., & Visscher, A. J. (2017). The effects of a digital formative assessment tool on mathematics achievement and student motivation: Results of a randomized experiment. *Computers & education*, 106, 83-96.
- Falchikov, N., & Goldfinch, J. (2000). Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks. *Review of educational research*, 70(3), 287-322.
- Fallon, M., & Forrest, S. L. (2011). High-tech versus low-tech instructional strategies: A comparison of clickers and handheld response cards. *Teaching of Psychology*, 38(3), 194-198.
- Fies, C., & Marshall, J. (2006). Classroom response systems: A review of the literature. *Journal of Science Education and Technology*, 15(1), 101-109.

- Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American psychologist*, 34(10), 906.
- Fortner-Wood, C., Armistead, L., Marchand, A., & Morris, F. B. (2013). The effects of student response systems on student learning and attitudes in undergraduate psychology courses. *Teaching of Psychology*, 40(1), 26-30.
- Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., & Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences*, 111(23), 8410-8415.
- Fulkerson, F. E., & Martin, G. (1981). Effects of exam frequency on academic performance, evaluations of instructor, and test anxiety. *Teaching of Psychology*, 8(2), 90-93.
- Fulmer, R. S. (1976). Test anxiety (worry and emotionality) changes during academic testing as a function of feedback and test importance. *Journal of educational psychology*, 68(6), 817-824.
- Furtak, E. M., Kiemer, K., Circi, R. K., Swanson, R., de León, V., Morrison, D., & Heredia, S. C. (2016). Teachers' formative assessment abilities and their relationship to student learning: findings from a four-year intervention study. *Instructional Science*, 44(3), 267-291.
- Gick, M. L. & Holyoak KJ (1983): Schema induction and analogical transfer. *Cognitive Psychology*, 15(1.38).
- Gielen, S., Tops, L., Dochy, F., Onghena, P., & Smeets, S. (2010). A comparative study of peer and teacher feedback and of various peer feedback forms in a secondary school writing curriculum. *British educational research journal*, 36(1), 143-162.
- Gipps, C. V. (2005). What is the role for ICT-based assessment in universities?. *Studies in Higher Education*, 30(2), 171-180.
- Glynn, S. M., Taasobshirazi, G., & Brickman, P. (2009). Science motivation questionnaire: Construct validation with nonscience majors. *Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching*, 46(2), 127-146.
- Glynn, S., & Koballa, M. (2006). *Motivation to Learn in College Science. Chpt. 3 in Mintzes, J. & Leonard, W* (Doctoral dissertation, ed. s), Handbook of College Science Teaching, Danvers, NSTA Press).
- Goetz, T., Bieg, M., Lüdtke, O., Pekrun, R., & Hall, N. C. (2013). Do girls really experience more anxiety in mathematics?. *Psychological Science*, 24(10), 2079-2087.
- González, A., Fernández, M. V. C., & Paoloni, P. V. (2017). Hope and anxiety in physics class: Exploring their motivational antecedents and influence on metacognition and performance. *Journal of Research in Science Teaching*, 54(5), 558-585.
- Griffin, T. D., Jee, B. D., & Wiley, J. (2009). The effects of domain knowledge on metacomprehension accuracy. *Memory & Cognition*, 37(7), 1001-1013.
- Grove, W. M., & Andreasen, N. C. (1982). Simultaneous tests of many hypotheses in exploratory research. *Journal of Nervous and Mental Disease.*, 170, 3, 8

## References

- Guarascio, A. J., Nemecek, B. D., & Zimmerman, D. E. (2017). Evaluation of students' perceptions of the Socrative application versus a traditional student response system and its impact on classroom engagement. *Currents in Pharmacy Teaching and Learning*, 9(5), 808-812.
- Guse, D. M., & Zobitz, P. M. (2011). Validation of the audience response system. *British Journal of Educational Technology*, 42(6), 985-991.
- Haelermans, C., & Ghysels, J. (2017). The effect of individualized digital practice at home on math skills—Evidence from a two-stage experiment on whether and why it works. *Computers & Education*, 113, 119-134.
- Hake, R. R. (1998). Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American journal of Physics*, 66(1), 64-74.
- Han, J. H., & Finkelstein, A. (2013). Understanding the effects of professors' pedagogical development with Clicker Assessment and Feedback technologies and the impact on students' engagement and learning in higher education. *Computers & Education*, 65, 64-76.
- Harrison, G. M., & Vallin, L. M. (2018). Evaluating the metacognitive awareness inventory using empirical factor-structure evidence. *Metacognition and Learning*, 13(1), 15-38.
- Hattie, J. A. C., & Learning, V. (2009). A synthesis of over 800 meta-analyses relating to achievement. New York.
- Hattie, J., & Gan, M. (2011). Instruction based on feedback. In P. Alexander & R. E. Mayer (Eds.), *Handbook of research on learning and instruction* (pp. 249–271). New York, NY: Routledge.
- Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of educational research*, 77(1), 81-112.
- Havnes, A., Smith, K., Dysthe, O., & Ludvigsen, K. (2012). Formative assessment and feedback: Making learning visible. *Studies in Educational Evaluation*, 38(1), 21-27.
- Herman, W. E. (1990). Fear of failure as a distinctive personality trait measure of test anxiety. *Journal of Research & Development in Education*.
- Hoekstra, A. (2008). Vibrant student voices: Exploring effects of the use of clickers in large college courses. *Learning, Media and Technology*, 33(4), 329-341.
- Hoekstra, A., & Mollborn, S. (2012). How clicker use facilitates existing pedagogical practices in higher education: data from interdisciplinary research on student response systems. *Learning, Media and Technology*, 37(3), 303-320.
- Hoffmann-Biencourt, A., Lockl, K., Schneider, W., Ackerman, R., & Koriat, A. (2010). Self-paced study time as a cue for recall predictions across school age. *British Journal of Developmental Psychology*, 28(4), 767-784.
- Hong, Z. R. (2010). Effects of a collaborative science intervention on high achieving students' learning anxiety and attitudes toward science. *International journal of science education*, 32(15), 1971-1988.

- Hox, J. (1998). Multilevel modeling: When and why. In *Classification, data analysis, and data highways* (pp. 147-154). Springer, Berlin, Heidelberg.
- Hubbard, J. K., & Couch, B. A. (2018). The positive effect of in-class clicker questions on later exams depends on initial academic performance level but not question format. *Computers & Education*, 120, 1-12.
- Hunsu, N. J., Adesope, O., & Bayly, D. J. (2016). A meta-analysis of the effects of audience response systems (clicker-based technologies) on cognition and affect. *Computers & Education*, 94, 102-119.
- Hwang, G. J., & Chang, H. F. (2011). A formative assessment-based mobile learning approach to improving the learning attitudes and achievements of students. *Computers & Education*, 56(4), 1023-1031.
- Hwang, G. J., & Tsai, C. C. (2011). Research trends in mobile and ubiquitous learning: A review of publications in selected journals from 2001 to 2010. *British Journal of Educational Technology*, 42(4), E65-E70.
- Isaacson, R., & Fujita, F. (2006). Metacognitive knowledge monitoring and self-regulated learning. *Journal of the Scholarship of Teaching and Learning*, 39-55.
- Iverson, A. M., Iverson, G. L., & Lukin, L. E. (1994). Frequent, ungraded testing as an instructional strategy. *The Journal of experimental education*, 62(2), 93-101.
- James, M. C., & Willoughby, S. (2011). Listening to student conversations during clicker questions: What you have not heard might surprise you! *American Journal of Physics*, 79(1), 123-132.
- Jones, D. (2007). Speaking, listening, planning and assessing: the teacher's role in developing metacognitive awareness. *Early Child Development and Care*, 177(6-7), 569-579.
- Jones, M. E., Antonenko, P. D., & Greenwood, C. M. (2012). The impact of collaborative and individualized student response system strategies on learner motivation, metacognition, and knowledge transfer. *Journal of Computer Assisted Learning*, 28(5), 477-487.
- Jonsson, A. (2013). Facilitating productive use of feedback in higher education. *Active learning in higher education*, 14(1), 63-76.
- Jovanovska, J. (2018). Designing effective multiple-choice questions for assessing learning outcomes. *INFOtheca-Journal for Digital Humanities*, 18(1), 25-42.
- Jurik, V., Gröschner, A., & Seidel, T. (2013). How student characteristics affect girls' and boys' verbal engagement in physics instruction. *Learning and instruction*, 23, 33-42.
- Juwah, C., Macfarlane-Dick, D., Matthew, B., Nicol, D., Ross, D., & Smith, B. (2004). Enhancing student learning through effective formative feedback. *The Higher Education Academy*, 140, 1-40.
- Kaddoura, M. (2013). Think pair share: A teaching learning strategy to enhance students' critical thinking. *Educational Research Quarterly*, 36(4), 3.
- Kang, H., Lundeberg, M., Wolter, B., delMas, R., & Herreid, C. F. (2012). Gender differences in student performance in large lecture classrooms using personal response systems ('clickers') with narrative case studies. *Learning, Media and Technology*, 37(1), 53-76.

## References

- Karasel, N., Ayda, O., & Tezer, M. (2010). The relationship between mathematics anxiety and mathematical problem solving skills among primary school students. *Procedia-Social and Behavioral Sciences*, 2(2), 5804-5807.
- Kay, R. H., & LeSage, A. (2009). Examining the benefits and challenges of using audience response systems: A review of the literature. *Computers & Education*, 53(3), 819-827.
- Kay, R., LeSage, A., & Knaack, L. (2010). Examining the use of audience response systems in secondary school classrooms: A formative analysis. *Journal of Interactive Learning Research*, 21(3), 343-365.
- Kelley, C. M., & Rhodes, M. G. (2002). Making sense and nonsense of experience: Attributions in memory and judgment. In *Psychology of learning and motivation* (Vol. 41, pp. 293-320). Academic Press.
- Keough, S. M. (2012). Clickers in the classroom: A review and a replication. *Journal of Management Education*, 36(6), 822-847.
- Ketabi, S., & Ketabi, S. (2014). Classroom and Formative Assessment in Second/Foreign Language Teaching and Learning. *Theory & Practice in Language Studies*, 4(2).
- Khan, R. A., & Jawaid, M. (2020). Technology enhanced assessment (TEA) in COVID 19 pandemic. *Pakistan journal of medical sciences*, 36(COVID19-S4), S108.
- King, D. B., & Joshi, S. (2008). Gender differences in the use and effectiveness of personal response devices. *Journal of Science Education and Technology*, 17(6), 544-552.
- Kjolsing, E., & Van Den Eide, L. (2016). Peer instruction: Using isomorphic questions to document learning gains in a small statics class. *Journal of Professional Issues in Engineering Education and Practice*, 142(4), 04016005.
- Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological bulletin*, 119(2), 254.
- Knight, J. K., & Brame, C. J. (2018). Peer instruction. *CBE—Life Sciences Education*, 17(2), fe5.
- Knight, J. K., Wise, S. B., & Southard, K. M. (2013). Understanding clicker discussions: student reasoning and the impact of instructional cues. *CBE—Life Sciences Education*, 12(4), 645-654.
- Knight, J. K., Wise, S. B., Rentsch, J., & Furtak, E. M. (2015). Cues matter: learning assistants influence introductory biology student interactions during clicker-question discussions. *CBE-Life Sciences Education*, 14(4), ar41.
- Knight, J. K., & Wood, W. B. (2005). Teaching more by lecturing less. *Cell biology education*, 4(4), 298-310.
- Koka, A., & Hein, V. (2006). Perceptions of teachers' positive feedback and perceived threat to sense of self in physical education: a longitudinal study. *European Physical Education Review*, 12(2), 165-179.
- Kokina, J., & Juras, P. E. (2017). Using Socratic to Enhance Instruction in an Accounting Classroom. *Journal of Emerging Technologies in Accounting*, 14(1), 85-97.
- Kollar, I., & Fischer, F. (2010). Peer assessment as collaborative learning: A cognitive perspective. *Learning and instruction*, 20(4), 344-348.

- 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., & Levy-Sadot, R. (2000). Conscious and unconscious metacognition: A rejoinder.
- Koriat, A., Nussinson, R., Bless, H., & Shaked, N. (2008). Information-based and experience-based metacognitive judgments: Evidence from subjective confidence. *A handbook of memory and metamemory*, 117-136.
- Kornell, N., & Son, L. K. (2009). Learners' choices and beliefs about self-testing. *Memory*, 17(5), 493-501.
- Koth, C. W., Bradshaw, C. P., & Leaf, P. J. (2008). A multilevel study of predictors of student perceptions of school climate: The effect of classroom-level factors. *Journal of Educational Psychology*, 100(1), 96.
- Krumsvik, R. (2012). Feedback Clickers in Plenary Lectures: A New Tool for Formative Assessment?. In *Transformative approaches to new technologies and student diversity in futures oriented classrooms* (pp. 191-216). Springer, Dordrecht.
- Krumsvik, R. J., & Ludvigsen, K. (2012). Formative E-assessment in plenary lectures. *Nordic Journal of Digital Literacy*, 7(01), 36-54.
- Langen, A. V., & Dekkers, H. (2005). Cross-national differences in participating in tertiary science, technology, engineering and mathematics education. *Comparative Education*, 41(3), 329-350.
- Lantz, M. E. (2010). The use of 'clickers' in the classroom: Teaching innovation or merely an amusing novelty?. *Computers in Human Behavior*, 26(4), 556-561.
- Lantz, M. E., & Stawiski, A. (2014). Effectiveness of clickers: Effect of feedback and the timing of questions on learning. *Computers in Human Behavior*, 31, 280-286.
- Larsen, D. P., & Butler, A. C. (2013). Test-enhancing learning. In K. Walsh (Ed.), *Oxford textbook of medical education* (pp. 443-452). Oxford: Oxford University Press.
- Lasry, N., Charles, E., Whittaker, C., & Lautman, M. (2009, November). When talking is better than staying quiet. In *AIP Conference Proceedings* (Vol. 1179, No. 1, pp. 181-184). AIP.
- Lasry, N., Mazur, E., & Watkins, J. (2008). Peer instruction: From Harvard to the two-year college. *American journal of Physics*, 76(11), 1066-1069.
- Laugksch, R. C. (2000). Scientific literacy: A conceptual overview. *Science Education*, 84, 71-94.
- Lee, S. C., IrvIng, K., Pape, S., & Owens, D. (2015). Teachers' use of interactive technology to enhance students' metacognition: Awareness of student learning and feedback. *Journal of Computers in Mathematics and Science Teaching*, 34(2), 175-198.
- Lee, A. M., Keh, N. C., & Magill, R. A. (1993). Instructional effects of teacher feedback in physical education. *Journal of Teaching in Physical Education*, 12(3), 228-243.
- Levesque, A. A. (2011). Using clickers to facilitate development of problem-solving skills. *CBE—Life Sciences Education*, 10(4), 406-417.
- Lewin, J. D., Vinson, E. L., Stetzer, M. R., & Smith, M. K. (2016). A campus-wide investigation of clicker implementation: the status of peer discussion in STEM classes. *CBE—Life Sciences Education*, 15(1), ar6.

## References

- Lin, C. H., & Huang, Y. (2018). Tell me only what I want to know: Congruent self-motivation and feedback. *Social Behavior and Personality: an international journal*, 46(9), 1523-1536.
- Lin, Y. C., Liu, T. C., & Chu, C. C. (2011). Implementing clickers to assist learning in science lectures: The Clicker-Assisted Conceptual Change model. *Australasian Journal of Educational Technology*, 27(6), 979-996.
- Lipsey, M. W., Puzio, K., Yun, C., Hebert, M. A., Steinka-Fry, K., Cole, M. W., ... & Busick, M. D. (2012). Translating the Statistical Representation of the Effects of Education Interventions into More Readily Interpretable Forms. *National Center for Special Education Research*.
- Lopez, J. A., Love, C., & Watters, D. (2014). Clickers in Biosciences: Do they Improve Academic Performance?. *International Journal of Innovation in Science and Mathematics Education (formerly CAL-laborate International)*, 22(3).
- Lorenzo, M., Crouch, C. H., & Mazur, E. (2006). Reducing the gender gap in the physics classroom. *American Journal of Physics*, 74(2), 118-122.
- Ludvigsen, K., Johan Krumsvik, R., & Breivik, J. (2020). Behind the scenes: Unpacking student discussion and critical reflection in lectures. *British Journal of Educational Technology*.
- Ludvigsen, K., Krumsvik, R., & Furnes, B. (2015). Creating formative feedback spaces in large lectures. *Computers & Education*, 88, 48-63.
- Lusk, S. L. (1981). Test anxiety, level and accuracy of predicted performance. *Psychological reports*, 49(2), 527-532.
- Lynch, R., McNamara, P. M., & Seery, N. (2012). Promoting deep learning in a teacher education programme through self-and peer-assessment and feedback. *European Journal of Teacher Education*, 35(2), 179-197.
- Mallow, J. V. (1986). *Science Anxiety: Fear of Science and How to Overcome It* (revised edition), H&H Publications, Clearwater, FL.
- Mallow, J. V. (2006). Science anxiety: Research and action. In J. J. Mintzes & W. H. Leonard (Eds.), *Handbook of college science teaching* (pp. 3–14). Arlington, VA: NSTA Press.
- Maloney, E. A., Schaeffer, M. W., & Beilock, S. L. (2013). Mathematics anxiety and stereotype threat: Shared mechanisms, negative consequences and promising interventions. *Research in Mathematics Education*, 15(2), 115-128.
- Marnell, G. (2012). The perils of popularising science. The Huffington Post.
- Marx, J. D., & Cummings, K. (2007). Normalized change. *American Journal of Physics*, 75(1), 87-91.
- Mathan, S. A., & Koedinger, K. R. (2005). Fostering the intelligent novice: Learning from errors with metacognitive tutoring. *Educational psychologist*, 40(4), 257-265.
- Maule, R. W. (2001). Framework for metacognitive mapping to design metadata for intelligent hypermedia presentations. *Journal of Educational Multimedia and Hypermedia*, 10(1), 27-45.

- Mayer, R. E., Stull, A., DeLeeuw, K., Almeroth, K., Bimber, B., Chun, D., ... & Zhang, H. (2009). Clickers in college classrooms: Fostering learning with questioning methods in large lecture classes. *Contemporary educational psychology*, 34(1), 51-57.
- Mazur, E. (1997). Peer instruction (pp. 9-18). Upper Saddle River, NJ: Prentice Hall.
- McCombs, B. L. (1996). Alternative perspectives for motivation. *Developing engaged readers in school and home communities*, 67-87.
- McDaniel, M. A., Agarwal, P. K., Huelser, B. J., McDermott, K. B., & Roediger III, H. L. (2011). Test-enhanced learning in a middle school science classroom: The effects of quiz frequency and placement. *Journal of Educational Psychology*, 103(2), 399, 414.
- McDaniel, M. A., Roediger, H. L., & McDermott, K. B. (2007). Generalizing test-enhanced learning from the laboratory to the classroom. *Psychonomic bulletin & review*, 14(2), 200-206.
- McDonough, K., & Foote, J. A. (2015). The impact of individual and shared clicker use on students' collaborative learning. *Computers & Education*, 86, 236-249.
- McNeish, D. M. (2014). Analyzing clustered data with OLS regression: The effect of a hierarchical data structure. *Multiple Linear Regression Viewpoints*, 40, 11-16.
- Miesner, M. T., & Maki, R. H. (2007). The role of test anxiety in absolute and relative metacomprehension accuracy. *European Journal of Cognitive Psychology*, 19(4-5), 650-670.
- Miller, K., Schell, J., Ho, A., Lukoff, B., & Mazur, E. (2015). Response switching and self-efficacy in Peer Instruction classrooms. *Physical Review Special Topics-Physics Education Research*, 11(1), 010104.
- Miller, T. (2009). Formative computer-based assessment in higher education: The effectiveness of feedback in supporting student learning. *Assessment & Evaluation in Higher Education*, 34(2), 181-192.
- Mitchum, A. L., & Kelley, C. M. (2010). Solve the problem first: Constructive solution strategies can influence the accuracy of retrospective confidence judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 36(3), 699.
- Molin, F., Cabus, S., Haelermans, C., & Groot, W. (2019). Toward Reducing Anxiety and Increasing Performance in Physics Education: Evidence from a Randomized Experiment. *Research in Science Education*, 51, 233-249.
- Molin, F., Haelermans, C., Cabus, S., & Groot, W. (2020). The effect of feedback on Metacognition - A Randomized Experiment using Polling Technology. *Computers & Education*, 103885.
- Molin, F., Haelermans, C., Cabus, S., & Groot, W. (2021). Do feedback strategies improve students' learning gain? - Results of a randomized experiment using polling technology in physics classrooms. *Computers & Education*, 175, 104339.
- Morling, B., McAuliffe, M., Cohen, L., & DiLorenzo, T. M. (2008). Efficacy of personal response systems ("clickers") in large, introductory psychology classes. *Teaching of Psychology*, 35(1), 45-50.



## References

- Morsanyi, K., Cheallaigh, N. N., & Ackerman, R. (2019). Mathematics anxiety and metacognitive processes: Proposal for a new line of inquiry. *Psihologijske teme*, 28(1), 147-169.
- Mostafa, T., Echazarra, A., & Guillou, H. (2018). The science of teaching science: An exploration of science teaching practices in PISA 2015.
- Mota, A. R., Körhasan, N. D., Miller, K., & Mazur, E. (2019). Homework as a metacognitive tool in an undergraduate physics course. *Physical Review Physics Education Research*, 15(1), 010136.
- Moulton, B. R. (1986). Random group effects and the precision of regression estimates. *Journal of econometrics*, 32(3), 385-397.
- Mullaney, K. M., Carpenter, S. K., Grotenhuis, C., & Burianek, S. (2014). Waiting for feedback helps if you want to know the answer: The role of curiosity in the delay-of-feedback benefit. *Memory & cognition*, 42(8), 1273-1284.
- Mullet, H. G., Butler, A. C., Verdin, B., von Borries, R., & Marsh, E. J. (2014). Delaying feedback promotes transfer of knowledge despite student preferences to receive feedback immediately. *Journal of Applied Research in Memory and Cognition*, 3, 222-229.
- Myers, S. J., Rhodes, M. G., & Hausman, H. E. (2020). Judgments of learning (JOLs) selectively improve memory depending on the type of test. *Memory & cognition*, 1-14.
- Narciss, S. (2008). Feedback strategies for interactive learning tasks. *Handbook of research on educational communications and technology*, 3, 125-144.
- Naveh-Benjamin, M. (1991). A comparison of training programs intended for different types of test-anxious students: Further support for an information-processing model. *Journal of Educational Psychology*, 83(1), 134.
- Naveh-Benjamin, M., McKeachie, W. J., Lin, Y. G., & Holinger, D. P. (1981). Test anxiety: deficits in information processing. *Journal of educational psychology*, 73(6), 816.
- Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and some new findings. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 26, pp. 125–173). New York: Academic Press.
- Neto, A., & Valente, M. O. (1997). Problem Solving in Physics: Towards a Metacognitively Developed Approach.
- Nicaise, V., Cogérino, G., Fairclough, S., Bois, J., & Davis, K. (2007). Teacher feedback and interactions in physical education: Effects of student gender and physical activities. *European Physical Education Review*, 13(3), 319-337.
- Nicol, D. J., & Boyle, J. T. (2003). Peer instruction versus class-wide discussion in large classes: A comparison of two interaction methods in the wired classroom. *Studies in higher education*, 28(4), 457-473.
- Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in higher education*, 31(2), 199-218.

- Nielsen, K. L., Hansen, G., & Stav, J. B. (2013). Teaching with student response systems (SRS): teacher-centric aspects that can negatively affect students' experience of using SRS. *Research in Learning Technology*, 21, 18989.
- Nielsen, K. L., Hansen-Nygård, G., & Stav, J. B. (2012). Investigating peer instruction: How the initial voting session affects students' experiences of group discussion. *International Scholarly Research Notices*, 2012.
- Noel, T. (2010). 2. Clickers 201: Exploring the Next Levels of Using Classroom Response Systems in Science Courses. *Collected Essays on Learning and Teaching*, 3, 9-14.
- Núñez-Peña, M. I., Bono, R., & Suárez-Pellicioni, M. (2015). Feedback on students' performance: A possible way of reducing the negative effect of math anxiety in higher education. *International Journal of Educational Research*, 70, 80-87.
- OECD (2005), *Formative Assessment: Improving Learning in Secondary Classrooms*, OECD Publishing, Paris, <https://doi.org/10.1787/9789264007413-en>.
- OECD (2007), *PISA 2006: Science Competencies for Tomorrow's World: Volume 1: Analysis*, PISA, OECD Publishing, Paris, <https://doi.org/10.1787/9789264040014-en>.
- OECD (2016), *PISA 2015 Results (Volume I): Excellence and Equity in Education*, PISA, OECD Publishing, Paris, <https://doi.org/10.1787/9789264266490-en>.
- OECD, *PISA 2015 Results in Focus*. PISA, OECD Publishing, 4-14 (2016).
- Oigara, J., & Keengwe, J. (2013). Students' perceptions of clickers as an instructional tool to promote active learning. *Education and Information Technologies*, 18(1), 15-28.
- Oswald, K. M., Blake, A. B., & Santiago, D. T. (2014). Enhancing immediate retention with clickers through individual response identification. *Applied Cognitive Psychology*, 28(3), 438-442.
- Özcan, Z. Ç. (2016). The relationship between mathematical problem-solving skills and self-regulated learning through homework behaviours, motivation, and metacognition. *International Journal of Mathematical Education in Science and Technology*, 47(3), 408-420.
- Pan, S. C., Cooke, J., Little, J. L., McDaniel, M. A., Foster, E. R., Connor, L. T., & Rickard, T. C. (2019). Online and clicker quizzing on jargon terms enhances definition-focused but not conceptually focused biology exam performance. *CBE—Life Sciences Education*, 18(4), ar54.
- Panaoura, A., & Philippou, G. (2005). The measurement of young pupils' metacognitive ability in mathematics: The case of self-representation and self-evaluation. In *Proceedings of CERME (Vol. 4)*.
- Paris, S. G., & Paris, A. H. (2001). Classroom applications of research on self-regulated learning. *Educational psychologist*, 36(2), 89-101.
- Pat-El, R., Tillema, H., & van Koppen, S. W. (2012). Effects of formative feedback on intrinsic motivation: Examining ethnic differences. *Learning and Individual Differences*, 22(4), 449-454.

## References

- Patterson, B., Kilpatrick, J., & Woebkenberg, E. (2010). Evidence for teaching practice: The impact of clickers in a large classroom environment. *Nurse education today*, 30(7), 603-607.
- Pekrun, R. (1992). The impact of emotions on learning and achievement: Towards a theory of cognitive/motivational mediators. *Applied Psychology*, 41(4), 359-376.
- Pekrun, R. (2006). The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice. *Educational psychology review*, 18(4), 315-341.
- Pekrun, R., Goetz, T., Daniels, L. M., Stupnisky, R. H., & Perry, R. P. (2010). Boredom in achievement settings: Exploring control-value antecedents and performance outcomes of a neglected emotion. *Journal of Educational Psychology*, 102(3), 531.
- Perez, K. E., Strauss, E. A., Downey, N., Galbraith, A., Jeanne, R., & Cooper, S. (2010). Does displaying the class results affect student discussion during peer instruction?. *CBE - Life Sciences Education*, 9(2), 133-140.
- Peugh, J. L. (2010). A practical guide to multilevel modeling. *Journal of school psychology*, 48(1), 85-112.
- Pintrich, P. R. (2000a). Issues in self-regulation theory and research. *The Journal of Mind and Behavior*, 213-219.
- Pintrich, P. R. (2000b). Multiple goals, multiple pathways: The role of goal orientation in learning and achievement. *Journal of educational psychology*, 92(3), 544.
- Pintrich, P. R., & De Groot, E. V. (1990). Motivational and self-regulated learning components of classroom academic performance. *Journal of educational psychology*, 82(1), 33.
- Pollock, S. J., Chasteen, S. V., Dubson, M., & Perkins, K. K. (2010, October). The use of concept tests and peer instruction in upper-division physics. In AIP conference proceedings (Vol. 1289, No. 1, pp. 261-264). *American Institute of Physics*.
- Porter, L., Bailey Lee, C., Simon, B., & Zingaro, D. (2011). Peer instruction: do students really learn from peer discussion in computing?. In *Proceedings of the seventh international workshop on Computing education research* (pp. 45-52).
- Premuroso, R. F., Tong, L., & Beed, T. K. (2011). Does using clickers in the classroom matter to student performance and satisfaction when taking the introductory financial accounting course?. *Issues in Accounting Education*, 26(4), 701-723.
- Pyc, M. A., Rawson, K. A., & Aschenbrenner, A. J. (2014). Metacognitive monitoring during criterion learning: When and why are judgments accurate?. *Memory & cognition*, 42(6), 886-897.
- Rakoczy, K., Klieme, E., Bürgermeister, A., & Harks, B. (2008). The interplay between student evaluation and instruction: Grading and feedback in mathematics classrooms. *Zeitschrift für Psychologie/Journal of Psychology*, 216(2), 111.
- Raudenbush, S. W. (1997). Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods*, 2(2), 173, 185.

- Rausch, J. R., Maxwell, S. E., & Kelley, K. (2003). Analytic Methods for Questions Pertaining to a Randomized Pretest, Posttest, Follow-up Design. *Journal of Clinical Child & Adolescent Psychology, 32*(3), 467-486.
- Reay, N. W., Li, P., & Bao, L. (2008). Testing a new voting machine question methodology. *American Journal of Physics, 76*(2), 171-178.
- Renner, C. H., & Renner, M. J. (2001). But I thought I knew that: Using confidence estimation as a debiasing technique to improve classroom performance. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition, 15*(1), 23-32.
- 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.
- Richardson, A. M., Dunn, P. K., McDonald, C., & Oprescu, F. (2015). CRISP: An instrument for assessing student perceptions of classroom response systems. *Journal of Science Education and Technology, 24*(4), 432-447.
- Rocklin, T., & Thompson, J. M. (1985). Interactive effects of test anxiety, test difficulty, and feedback. *Journal of Educational Psychology, 77*(3), 368, 372.
- Roediger III, H. L., & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science, 1*(3), 181-210.
- Roediger, H., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Science, 15*(1), 20-27.
- Roelle, J., Nowitzki, C., & Berthold, K. (2017). Do cognitive and metacognitive processes set the stage for each other?. *Learning and Instruction, 50*, 54-64.
- Ross, M. E., Green, S. B., Salisbury-Glennon, J. D., & Tollefson, N. (2006). College students' study strategies as a function of testing: An investigation into metacognitive self-regulation. *Innovative Higher Education, 30*(5), 361-375.
- Rothman, D. K. (2004). New approach to test anxiety. *Journal of College Student Psychotherapy, 18*(4), 45-60.
- Rozencajg, P. (2003). Metacognitive factors in scientific problem-solving strategies. *European Journal of Psychology of Education, 18*(3), 281-294.
- Ryan, R. M., & Deci, E. L. (2000). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary educational psychology, 25*(1), 54-67.
- Sadler, D. R. (1998). Formative assessment: Revisiting the territory. *Assessment in education: principles, policy & practice, 5*(1), 77-84.
- Sansone, C., & Thoman, D. B. (2005). Interest as the missing motivator in self-regulation. *European Psychologist, 10*(3), 175-186.
- Schellings, G., & Van Hout-Wolters, B. (2011). Measuring strategy use with self-report instruments: theoretical and empirical considerations. *Metacognition and Learning, 6*(2), 83-90.
- Schraw, G. J., & Impara, J. C. (2000). *Issues in the measurement of metacognition*. Buros Inst of Mental.

## References

- Schraw, G., & Dennison, R. S. (1994). Assessing metacognitive awareness. *Contemporary educational psychology*, 19(4), 460-475.
- Schraw, G., Olafson, L., Weibel, M., & Sewing, D. (2012). Metacognitive knowledge and field-based science learning in an outdoor environmental education program. *In Metacognition in science education* (pp. 57-77). Springer, Dordrecht.
- Schunk, D. H., & Zimmerman, B. (2009). *Motivation and self-regulated learning: theory, research, and applications*. New York: Routledge.
- Schunk, D. H., & Zimmerman, B. J. (Eds.). (1998). *Self-regulated learning: From teaching to self-reflective practice*. Guilford Press.
- Seymour, E., & Hewitt, N. M. (1997). *Talking about leaving* (p. 134). Westview Press, Boulder, CO.
- Shaffer, D. M., & Collura, M. J. (2009). Evaluating the effectiveness of a personal response system in the classroom. *Teaching of Psychology*, 36(4), 273-277.
- Shapiro, A. M., & Gordon, L. T. (2012). A controlled study of clicker-assisted memory enhancement in college classrooms. *Applied Cognitive Psychology*, 26(4), 635-643.
- Shapiro, A. M., Sims-Knight, J., O'Rielly, G. V., Capaldo, P., Pedlow, T., Gordon, L., & Monteiro, K. (2017). Clickers can promote fact retention but impede conceptual understanding: The effect of the interaction between clicker use and pedagogy on learning. *Computers & Education*, 111, 44.
- Shirley, M. L., & Irving, K. E. (2015). Connected classroom technology facilitates multiple components of formative assessment practice. *Journal of Science Education and Technology*, 24(1), 56-68.
- Shute, V. J. (2008). Focus on formative feedback. *Review of educational research*, 78(1), 153-189.
- Singh, C. (2005). Impact of peer interaction on conceptual test performance. *American journal of physics*, 73(5), 446-451.
- Singh, C. (2008). Assessing student expertise in introductory physics with isomorphic problems. II. Effect of some potential factors on problem solving and transfer. *Physical Review Special Topics-Physics Education Research*, 4(1), 010105.
- Smart, D. T., Kelley, C. A., & Conant, J. S. (1999). Marketing education in the year 2000: Changes observed and challenges anticipated. *Journal of Marketing Education*, 21(3), 206-216.
- Smith, H., Cooper, A., & Lancaster, L. (2002). Improving the quality of undergraduate peer assessment: A case for student and staff development. *Innovations in education and teaching international*, 39(1), 71-81.
- Smith, M. K., Wood, W. B., Adams, W. K., Wieman, C., Knight, J. K., Guild, N., & Su, T. T. (2009). Why peer discussion improves student performance on in-class concept questions. *Science*, 323(5910), 122-124.
- Smith, M. K., Wood, W. B., Krauter, K., & Knight, J. K. (2011). Combining peer discussion with instructor explanation increases student learning from in-class concept questions. *CBE—Life Sciences Education*, 10(1), 55-63.

- Snow, R. E. (1968). Brunswikian approaches to research on teaching. *American Educational Research Journal*, 5(4), 475-489.
- Sobel, M. E. (1982). Asymptotic confidence intervals for indirect effects in structural equation models. *Sociological methodology*, 13, 290-312.
- Spielberger, C. D., & Vagg, P. R. (Eds.). (1995). *Test anxiety: Theory, assessment, and treatment*. Taylor & Francis.
- Stiggins, R., & Chappuis, J. (2006). What a difference a word makes. *Journal of Staff Development*, 27(1), 10-14.
- Stowell, J. R., & Nelson, J. M. (2007). Benefits of electronic audience response systems on student participant, learning, and emotion. *Teaching of Psychology*, 34, 253-258.
- Sullivan, D. (2017). Mediating test anxiety through the testing effect in asynchronous, objective, online assessments at the university level. *Journal of Education and Training*, 4(2), 107-123.
- Sun, J. C. Y. (2014). Influence of polling technologies on student engagement: An analysis of student motivation, academic performance, and brainwave data. *Computers & Education*, 72, 80-89.
- Sun, J. C. Y., & Hsieh, P. H. (2018). Application of a Gamified Interactive Response System to Enhance the Intrinsic and Extrinsic Motivation, Student Engagement, and Attention of English Learners. *Journal of Educational Technology & Society*, 21(3), 104-116.
- Sung, Y. T., Chang, K. E., & Liu, T. C. (2016). The effects of integrating mobile devices with teaching and learning on students' learning performance: A meta-analysis and research synthesis. *Computers & Education*, 94, 252-275.
- Sungur, S., & Senler, B. (2009). An analysis of Turkish high school students' metacognition and motivation. *Educational research and evaluation*, 15(1), 45-62.
- Sutherlin, A. L., Sutherlin, G. R., & Akpanudo, U. M. (2013). The effect of clickers in university science courses. *Journal of Science Education and Technology*, 22(5), 651-666.
- Swanson, H. L. (1990). Influence of metacognitive knowledge and aptitude on problem solving. *Journal of educational psychology*, 82(2), 306.
- Taasoobshirazi, G., Bailey, M., & Farley, J. (2015). Physics metacognition inventory part II: confirmatory factor analysis and rasch analysis. *International Journal of Science Education*, 37(17), 2769-2786.
- Tanner, K. D. (2009). Talking to learn: why biology students should be talking in classrooms and how to make it happen. *CBE—Life Sciences Education*, 8(2), 89-94.
- Tartavulea, C. V., Albu, C. N., Albu, N., Dieaconescu, R. I., & Petre, S. (2020). Online Teaching Practices and the Effectiveness of the Educational Process in the Wake of the COVID-19 Pandemic. *Amfiteatru Economic*, 22(55), 920-936.
- Tauber, S. K., Witherby, A. E., & Dunlosky, J. (2019). Beliefs about memory decline in aging do not impact judgments of learning (JOLs): A challenge for belief-based explanations of JOLs. *Memory & cognition*, 47(6), 1102-1119.
- Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International journal of medical education*, 2, 53.

## References

- Theobald, R., & Freeman, S. (2014). Is it the intervention or the students? Using linear regression to control for student characteristics in undergraduate STEM education research. *CBE—Life Sciences Education*, 13(1), 41-48.
- Thiede, K. W., & Dunlosky, J. (1999). Toward a general model of self-regulated study: An analysis of selection of items for study and self-paced study time. *Journal of experimental psychology: Learning, Memory, and Cognition*, 25(4), 1024.
- Thiede, K. W., Griffin, T. D., Wiley, J., & Anderson, M. C. (2010). Poor metacomprehension accuracy as a result of inappropriate cue use. *Discourse Processes*, 47(4), 331-362.
- Thiede, K. W., Redford, J. S., Wiley, J., & Griffin, T. D. (2017). How restudy decisions affect overall comprehension for seventh-grade students. *British Journal of Educational Psychology*, 87(4), 590-605.
- Thomas, G. P. (2013). Changing the metacognitive orientation of a classroom environment to stimulate metacognitive reflection regarding the nature of physics learning. *International Journal of Science Education*, 35(7), 1183-1207.
- Tlhoale, M., Hofman, A., Winnips, K., & Beetsma, Y. (2014). The impact of interactive engagement methods on students' academic achievement. *Higher Education Research & Development*, 33(5), 1020-1034.
- Tobias, S. (1985). Test anxiety: Interference, defective skills, and cognitive capacity. *Educational Psychologist*, 20(3), 135-142.
- Tobias, S., & Everson, H. T. (1997). Studying the relationship between affective and metacognitive variables. *Anxiety, Stress, and Coping*, 10(1), 59-81.
- Topping, K. (1998). Peer assessment between students in colleges and universities. *Review of educational Research*, 68(3), 249-276.
- Topping, K. J., Smith, E. F., Swanson, I., & Elliot, A. (2000). Formative peer assessment of academic writing between postgraduate students. *Assessment & evaluation in higher education*, 25(2), 149-169.
- Trees, A. R., & Jackson, M. H. (2007). The learning environment in clicker classrooms: student processes of learning and involvement in large university-level courses using student response systems. *Learning, Media and Technology*, 32(1), 21-40.
- Tsai, C. C., Lin, S. S., & Yuan, S. M. (2002). Developing science activities through a networked peer assessment system. *Computers & Education*, 38(1-3), 241-252.
- Tullis, J. G., & Goldstone, R. L. (2020). Why does peer instruction benefit student learning?. *Cognitive Research: Principles and Implications*, 5, 1-12.
- Udo, M. K., Ramsey, G. P., & Mallow, J. V. (2004). Science anxiety and gender in students taking general education science courses. *Journal of Science Education and Technology*, 13(4), 435-446.
- Udo, M. K., Ramsey, G. P., Reynolds-Alpert, S., & Mallow, J. V. (2001). Does physics teaching affect gender-based science anxiety?. *Journal of Science Education and Technology*, 10(3), 237-247.

- Ulbig, S. G., & Notman, F. (2012). Is class appreciation just a click away?: Using student response system technology to enhance shy students' introductory American government experience. *Journal of Political Science Education*, 8(4), 352-371.
- Van der Kleij, F. M., Feskens, R. C., & Eggen, T. J. (2015). Effects of feedback in a computer-based learning environment on students' learning outcomes: A meta-analysis. *Review of educational research*, 85(4), 475-511.
- Van der Klink, M. (2011). Bekwaam innoveren voor een toekomstbestendig hoger beroepsonderwijs.
- Van Loon, M. H., de Bruin, A. B., van Gog, T., van Merriënboer, J. J., & Dunlosky, J. (2014). Can students evaluate their understanding of cause-and-effect relations? The effects of diagram completion on monitoring accuracy. *Acta psychologica*, 151, 143-154.
- Van Merriënboer, J. J., & de Bruin, A. B. (2019). Cue-based facilitation of self-regulated learning: A discussion of multidisciplinary innovations and technologies. *Computers in Human Behavior*, 100, 384-391.
- Veenman, M. V., Kerseboom, L., & Imthorn, C. (2000). Test anxiety and metacognitive skillfulness: Availability versus production deficiencies. *Anxiety, Stress and Coping*, 13(4), 391-412.
- Versteeg, M., van Blankenstein, F. M., Putter, H., & Steendijk, P. (2019). Peer instruction improves comprehension and transfer of physiological concepts: A randomized comparison with self-explanation. *Advances in Health Sciences Education*, 24(1), 151-165.
- Vickrey, T., Rosploch, K., Rahmanian, R., Pilarz, M., & Stains, M. (2015). based implementation of peer instruction: A literature review. *CBE—Life Sciences Education*, 14(1), es3.
- Vital, F. (2011). Creating a positive learning environment with the use of clickers in a high school chemistry classroom. *Journal of Chemical Education*, 89(4), 470-473.
- Voerman, L., Meijer, P. C., Korthagen, F. A., & Simons, R. J. (2012). Types and frequencies of feedback interventions in classroom interaction in secondary education. *Teaching and Teacher Education*, 28(8), 1107-1115.
- Vollmeyer, R., & Rheinberg, F. (1999). Motivation and metacognition when learning a complex system. *European Journal of Psychology of Education*, 14(4), 541-554.
- Wang, A. I. (2015). The wear out effect of a game-based student response system. *Computers & Education*, 82, 217-227.
- White, B. Y., & Frederiksen, J. R. (1998). Inquiry, modeling, and metacognition: Making science accessible to all students. *Cognition and instruction*, 16(1), 3-118.
- White, B., & Frederiksen, J. (2005). A theoretical framework and approach for fostering metacognitive development. *Educational Psychologist*, 40(4), 211-223.
- Wiggs, C. M. (2011). Collaborative testing: Assessing teamwork and critical thinking behaviors in baccalaureate nursing students. *Nurse Education Today*, 31(3), 279-282.
- Wiley, J., Griffin, T. D., & Thiede, K. W. (2005). Putting the comprehension in metacomprehension. *The Journal of General Psychology*, 132(4), 408-428.



## References

- Willermark, S. (2021). Who's There? Characterizing Interaction in Virtual Classrooms. *Journal of Educational Computing Research*, 0735633120988530.
- Williams, J. (2003). *The Skills for Life survey: A national needs and impact survey of literacy, numeracy and ICT skills* (No. 490). The Stationery Office.
- Winne, P. H., & Hadwin, A. F. (1998). Studying as self-regulated learning. *Metacognition in educational theory and practice*, 93, 27-30.
- Winne, P. H., & Hadwin, A. F. (2008). The weave of motivation and self-regulated learning. *Motivation and self-regulated learning: Theory, research, and applications*, 2, 297-314.
- Wolfe, A. (2012). Implementing collaborative learning methods in the political science classroom. *Journal of Political Science Education*, 8(4), 420-432.
- Wong, G. K. W., & Yang, M. (2017). Using ICT to facilitate instant and asynchronous feedback for students' learning engagement and improvements. In *Emerging practices in scholarship of learning and teaching in a digital era* (pp. 289-309). Springer, Singapore.
- Wooldridge, J. (2010). *Econometric analysis of cross-section and panel data* (2nd ed.). Cambridge, MA: MIT Press.
- Wright, D., Clark, J., & Tiplady, L. (2018). Designing for formative assessment: A toolkit for teachers. In *Classroom Assessment in Mathematics* (pp. 207-228). Springer, Cham.
- Yourstone, S. A., Kraye, H. S., & Albaum, G. (2008). Classroom questioning with immediate electronic response: Do clickers improve learning?. *Decision Sciences Journal of Innovative Education*, 6(1), 75-88.
- Yu, Z., Chen, W., Kong, Y., Sun, X. L., & Zheng, J. (2014). The impact of clickers instruction on cognitive loads and listening and speaking skills in college English class. *PloS one*, 9(9), e106626.
- Zemira, M., & Bracha, K. (2014). *Educational Research and Innovation Critical Maths for Innovative Societies The Role of Metacognitive Pedagogies: The Role of Metacognitive Pedagogies*. OECD publishing.
- Zhonggen, Y. (2017). The influence of clickers use on metacognition and learning outcomes in college English classroom. In *Exploring the New Era of Technology-Infused Education* (pp. 158-171). IGI Global.
- Zimmerman, B. J. (1989). A social cognitive view of self-regulated academic learning. *Journal of educational psychology*, 81(3), 329.
- Zimmerman, B. J. (2000). Attaining self-regulation: A social cognitive perspective. In *Handbook of self-regulation* (pp. 13-39). Academic Press.
- Zimmerman, B. J., & Labuhn, A. S. (2012). Self-regulation of learning: Process approaches to personal development.
- Zingaro, D., & Porter, L. (2014). Peer Instruction in computing: The value of instructor intervention. *Computers & Education*, 71, 87-96.





## Summary

## Summary

### Summary

Secondary education plays an important role in students' choice of science-related studies. Worldwide, a large number of students do not choose science studies because they have doubts about their science abilities. To increase the number of enrolments in science studies, education must give students the confidence that they are capable of understanding science. To achieve this goal, science teachers should provide immediate meaningful feedback more frequently in their classes, because feedback has a powerful influence on students' understanding (Black & Wiliam, 1998a; Hattie & Timperley, 2007). Providing immediate feedback can be organized by 'formative assessments', which is a general term for all those activities undertaken by teachers that provide information about students' understanding and are used to modify teaching and learning activities. However, formative assessments still prove difficult to implement in classrooms. Teachers often do not have enough time to provide feedback, or do not know which students understand the course content and which students do not. To address these problems, teachers may choose to use real-time data collected by student response systems (SRS). These systems allow students to answer multiple-choice questions posed by their teacher in front of the class. Students answer the questions using their SRS, such as clickers or smartphones with a web-based program (e.g. *Socrative* or *Kahoot!*). The answers are collected and depending on the teacher's use and purpose, the number of correct and incorrect answers are summarized and displayed as a histogram in front of the class. The efficient collecting of answers gives teachers the opportunity to provide feedback. They can use it to encourage students to clarify or improve their explanations, or to eliminate incorrect thinking. The way SRS are used depends on how teachers organize a formative assessment. They may choose an individual mode where students individually think through and answer questions, or a more active mode where students discuss their thinking with their peers before answering the question individually. A lot of research has been conducted on performance outcomes when students are formatively assessed with SRS. While research provided important insights, there is still a lack of evidence as to why performance improvements occur.

*Chapter 2* presents an experimental study that examines the causal relationship between formative assessments and academic performance on the one hand, and formative

assessments and anxiety in physics on the other hand. One hundred and thirty-nine physics students of one upper secondary school participated in the study. Seventy-three treated students used SRS, and sixty-six untreated students did not use in-class questioning or SRS, but instead followed traditional physics teaching. The only difference between the two conditions was the way the teacher interacted with the students at the end of a class. Three times a week, students in the treated condition were formatively assessed with SRS for approximately 10 to 15 minutes. Students were given limited time to individually think about or discuss a question with their peers, after which they answered the question individually with their SRS. After each answered question, a distribution of responses was projected and the teacher explained the correct and incorrect answers. This study shows that formative assessments improve academic performance in physics compared to traditional teaching, as treated students score significantly higher in a final exam than untreated students. This corresponds to a medium effect. A significant effect of formative testing compared to traditional teaching has also been shown for anxiety in physics. This is a medium to small effect. Chapter 2 also shows that anxiety fully mediates the effects of formative assessments on academic performance, meaning that these assessments reduce anxiety and in turn significantly affect academic performance. The findings in Chapter 2 confirm that anxiety in physics decreases when students become familiar with exam preparation activities, and when they answer questions anonymously without being graded or judged on their responses. Simultaneously, the formative assessments divide the course content into smaller units, allowing students to receive immediate feedback, which improves their understanding of the course content and helps them to achieve a higher grade on an exam. This is reinforced by the mediating effect of anxiety in physics, as anxiety in physics affects student well-being, and thus indirectly academic performance.

*Chapter 3* provides insight into feedback strategies during formative assessments with SRS. The study examines whether feedback strategies, such as teacher feedback or peer discussions combined with teacher feedback, affect students' comprehension when they answer multiple-choice questions. To this end, an experimental study was conducted with three conditions in upper secondary school physics. Each week, students answered conceptual multiple-choice questions with the web-based SRS *Socrative*. After answering a conceptual question, students in the individual condition received teacher feedback, while

## Summary

students in the cooperative condition discussed the concept question again with their peers and answered it for a second time before receiving feedback from the teacher. Students in the control condition did not receive feedback on concept questions, but only heard the correct answer from the teacher. Each concept question was followed by a paired isomorphic (similar) question. This study shows that students who receive teacher feedback, or have peer discussions combined with teacher feedback, achieve significantly more learning gains when answering an isomorphic question compared to students who receive no feedback. These students apply their acquired knowledge to new isomorphic questions. Most learning gains occur when peer discussions are combined with teacher feedback. The findings indicate that students who might learn from peer discussions (because they initially answered the concept question incorrectly, but corrected it after peer discussions), do indeed learn from peer discussions. These students answer proportionally more isomorphic questions correctly than students who receive no feedback or do not discuss with their peers, but who answer the concept question correctly. Students can share and (re)construct information in peer discussions while building on their own concepts or on the ideas of peers, leading to an additional reinforcement of understanding and transfer of acquired knowledge to new follow-up isomorphic questions.

*Chapter 4* presents an experimental study that examines the effects of both teacher feedback, and peer discussions combined with teacher feedback on metacognitive skills and motivation. Over a period of ten weeks, six hundred and thirty-three physics students from six secondary schools answered each week four paired sets of conceptual multiple-choice questions using the web-based SRS *Socratic*. Students in the individual condition and in the cooperative condition answered the multiple-choice questions individually. However, students in the cooperative condition discussed their responses with their peers before the answer was explained by the teacher. Students in the control condition heard only the correct answer, but did not discuss their responses in pairs, and received no teacher feedback. This study shows that peer discussions combined with teacher feedback have a significant effect on both metacognition and motivation. The effect is partly mediated by motivation. An explanation for this mediation effect is that peer discussions make formative assessments more interactive and interesting. Students become more proactive and focused when answering multiple-choice questions, which leads to an increased motivation and, indirectly, to more

metacognitive skills. Furthermore, the effects differ between girls and boys, as girls increase their metacognitive skills and get more motivated through teacher feedback (whether or not in combination with peer discussions) than boys, while boys increase their metacognitive skills and motivation only through peer discussions in combination with teacher feedback. Girls have a more positive attitude to teacher feedback and are more aware of the quality of this feedback, while boys improve their metacognitive skills when they are encouraged to discuss their thoughts. Chapter 4 also shows differential effects between students with different levels of metacognitive skills. Students with low metacognitive skills benefit more from peer discussions on top of teacher feedback than students with high metacognitive skills.

Many teachers use SRS in their formative assessments without realizing the extent to which these assessments affect student learning. Because formative assessments with SRS have not only direct learning effects but also metacognitive effects, *Chapter 5* provides insights into relationships of factors that may affect metacognition. The study describes the cue utilization framework of Koriat (1997) and the control model of Nelson and Narens (1990) and complements it with aspects of formative assessments when using SRS in a developed framework. The study shows that formative assessments contain multiple prompts, which are activities or activators that help students to engage in their cognitive processes. The most common prompts in formative assessments with SRS are *anonymous polling, histogram feedback, teacher feedback, peer discussions and isomorphic questions*. These prompts trigger previous knowledge of students (one more than the other) and help them to identify one or more diagnostic cues that are predictive of subsequent learning and performance. Cues are information that shift students' attention to the required knowledge. Common cues in formative assessments are *comprehension, course demands, motivation to learn and anxiety*. The framework argues that prompts during an assessment lead to diagnostic cues, that improve the accuracy of monitoring judgments. Formative assessments can be conducted in a variety of ways; a teacher can choose to show histograms or not, provide feedback or not, allow peer discussions or not, ask isomorphic questions or not, etc. So the number and variety of prompts students receive depends on how a teacher organizes a formative assessment. This study states that more prompts during an assessment lead to more diagnostic cues, resulting in an improvement of accuracy of monitoring judgments and an improvement of



## Summary

metacognitive skills. When teachers use these prompts together, they might be even more effective because they reinforce cues when judging monitoring and regulation of learning.

In *Chapter 6*, the findings and conclusions from previous chapters are compiled into five key conclusions. The first key conclusion is that *'providing feedback in formative assessments increases both short-term and long-term course understanding'*. Students who receive teacher feedback on concept questions, whether or not combined with peer discussions, increase their course understanding in the short term. The results imply effect sizes of  $d = 0.83$  for teacher feedback and  $d = 1.13$  for peer discussion combined with teacher feedback. In addition, students who are daily formatively assessed and receive feedback on their answers gain more course understanding in the long-term than students who do not receive these assessments and feedback on questions. This long-term course understanding corresponds to an effect size of  $d = 0.34$ . The second key conclusion is that *'students learn from peer discussion and thereby improve their metacognitive awareness'*. Students learn from peer discussions and do not simply copy the same answer of their more skilled peer. Peer discussions allow students to practice explaining concepts to one another, and explore answers or solutions they might not find on their own. Peer discussion combined with teacher feedback has also a positive effect on metacognition compared to students who receive no feedback. The third key conclusion is that *'the influence of feedback on metacognition differs by gender and level of metacognitive skills'*. First, there are different effects among boys and girls. Compared to boys, metacognitive skills increase for girls via both teacher feedback or a combination of peer discussions and teacher feedback, while for boys metacognitive skills increase only via peer discussions combined with teacher feedback. Second, there are differential effects among subgroups of students with different metacognitive skills. Low-metacognitive students benefit significantly more from peer discussions on top of teacher feedback than high-metacognitive students. The fourth key conclusion is that *'more course understanding leads to more metacognitive awareness when students discuss with their peers and receive feedback from their teachers'*. Course understanding after only passively listening to the teachers' feedback is sufficient to understand course content in the short term, but is insufficient to monitor and regulate learning in a way that it increases metacognitive skills. It is in fact the combination of teacher feedback with peer discussions that reinforce conceptual understanding and increase metacognitive awareness, as peer

discussions create an environment in which students demonstrate, explain, discuss, and control their own thought processes. The fifth key conclusion is that '*student feelings of anxiety in physics and motivation to learn mediate the effects of formative assessments*'. Feelings of anxiety in physics and feelings of motivation to learn mediate the effects of prompts on academic performance and metacognition, respectively. This mediation means that prompts trigger cues like anxiety cues and motivation cues, which in turn influence metacognition and academic performance.



## Samenvatting

## Samenvatting

De vakken natuurkunde, scheikunde, wiskunde en biologie spelen een belangrijke rol in de keuze van leerlingen voor een vervolgstudie in de exacte vakken. Wereldwijd kiest een groot aantal leerlingen niet voor een bètastudie omdat zij twijfelen aan hun capaciteiten. Om het aantal inschrijvingen voor bètastudies te verhogen, moet het onderwijs leerlingen het vertrouwen geven dat zij in staat zijn de exacte vakken te begrijpen. Om dit doel te bereiken, zouden docenten in bètavakken vaker onmiddellijke relevante feedback moeten geven in hun lessen, omdat feedback een grote invloed heeft op het begrip van leerlingen (Black & Wiliam, 1998a; Hattie & Timperley, 2007). Het geven van directe feedback kan georganiseerd worden door formatieve toetsen, wat een algemene term is voor al die activiteiten van docenten die informatie geven over het begrip van leerlingen en die gebruikt worden om onderwijs- en leeractiviteiten aan te passen. Toch blijkt dat formatief toetsen vaak nog moeilijk te implementeren is in lessen. Docenten hebben vaak niet genoeg tijd om feedback te geven, of weten niet welke leerlingen de lesstof wel begrijpen en welke leerlingen niet. Om deze knelpunten aan te pakken, kunnen docenten ervoor kiezen om real-time gegevens te gebruiken die worden verzameld door *student response systems*. Deze systemen stellen leerlingen in staat om meerkeuzevragen te beantwoorden die door hun docent worden gesteld. Leerlingen beantwoorden de vragen met behulp van hun eigen *student response system* (SRS), zoals clickers of smartphones met een webgebaseerd programma (bv. *Socrative* of *Kahoot!*). De antwoorden worden verzameld en afhankelijk van het gebruik en doel van de docent wordt het aantal goede en foute antwoorden samengevat en weergegeven als een histogram voor in de klas. Het efficiënt verzamelen van antwoorden geeft docenten de kans om onmiddellijke feedback te geven. Ze kunnen het gebruiken om leerlingen aan te moedigen hun uitleg te verduidelijken of te verbeteren, of om foute denkwijzen te corrigeren. De manier waarop SRS worden gebruikt hangt af van hoe docenten een formatieve toets organiseren. Ze kunnen kiezen voor een individuele aanpak waarin leerlingen individueel nadenken en vragen beantwoorden, of voor een meer coöperatieve aanpak waarin leerlingen hun denkwijzen bespreken met hun medeleerlingen alvorens de vraag individueel te beantwoorden. Er is veel onderzoek gedaan naar leerresultaten van leerlingen wanneer zij formatief getoetst worden met SRS. Hoewel deze onderzoeken belangrijke inzichten geven, is er nog steeds een gebrek aan bewijs voor de vraag waarom leerprestaties kunnen verbeteren.

*Hoofdstuk 2* presenteert de resultaten van een experimentele studie naar de causale relatie tussen formatief toetsen met SRS en leerprestaties aan de ene kant, en formatief toetsen met SRS en angst voor natuurkunde aan de andere kant. Honderdnegenendertig natuurkunde leerlingen van de bovenbouw van één middelbare school namen deel aan deze studie. Drieënzeventig leerlingen in de interventiegroep maakten gebruik van SRS, en zesenzestig leerlingen in de controlegroep maakten geen gebruik van meerkeuzevragen in de klas of van SRS, maar kregen in plaats daarvan een standaard natuurkunde les. Het enige verschil tussen de twee groepen was de manier waarop de docent met de leerlingen werkte aan het eind van de les. Drie keer per week werden de leerlingen in de interventiegroep gedurende 10 tot 15 minuten formatief getoetst met SRS. Leerlingen kregen kort de tijd om individueel na te denken over een vraag of deze te bespreken met hun klasgenoten, waarna ze de vraag individueel beantwoordden met hun SRS. Na elke beantwoorde vraag werd een histogram geprojecteerd en legde de docent de goede en foute antwoorden uit. Deze studie toont aan dat formatief toetsen de leerprestaties bij natuurkunde verbetert ten opzichte van leerlingen in de controlegroep, aangezien leerlingen in de interventiegroep significant hoger scoren op een toets dan leerlingen in de controlegroep. Dit komt overeen met een gemiddeld effect. Een significant effect van formatief toetsen ten opzichte van leerlingen in de controlegroep is ook aangetoond voor angst voor natuurkunde. Dit is een klein tot middelgroot effect. *Hoofdstuk 2* laat ook zien dat angst de effecten van formatieve toetsen op leerprestaties medieert, wat betekent dat deze toetsen angst verminderen en op hun beurt de leerprestaties significant beïnvloeden. De bevindingen in *Hoofdstuk 2* tonen aan dat angst voor natuurkunde afneemt wanneer leerlingen vertrouwd raken met toets voorbereidende activiteiten, en wanneer ze anoniem vragen beantwoorden zonder een cijfer te krijgen of beoordeeld te worden op hun antwoorden. Tegelijkertijd verdelen de formatieve toetsen de lesstof in kleinere porties, waardoor leerlingen onmiddellijk feedback krijgen, wat hun begrip van de leerstof verbetert en hen helpt om een hoger cijfer te halen op een toets. Dit wordt versterkt door het mediërende effect van angst voor natuurkunde, aangezien formatieve toetsingen het welzijn van leerlingen beïnvloeden, en dus ook indirect de leerprestaties.

*Hoofdstuk 3* geeft inzicht in feedbackstrategieën tijdens formatieve toetsingen met SRS. Onderzocht wordt of feedbackstrategieën, zoals feedback van de docent of peer-discussies in combinatie met feedback van de docent, van invloed zijn op het begrip van leerlingen tijdens

## Samenvatting

het beantwoorden van meerkeuzevragen. Hiertoe werd een experimentele studie uitgevoerd in de bovenbouw van het voortgezet onderwijs met drie verschillende natuurkunde groepen. Elke week beantwoordden leerlingen conceptuele meerkeuzevragen met het webgebaseerde SRS *Socratic*. Na het beantwoorden van een conceptuele vraag kregen leerlingen in de individuele groep feedback van de docent, terwijl leerlingen in de coöperatieve groep de conceptuele vraag opnieuw bespraken met hun medeleerlingen en de vraag een tweede keer beantwoordden voordat ze feedback kregen van de docent. Leerlingen in de controlegroep kregen geen feedback op de conceptvragen, maar hoorden alleen het correcte antwoord van de docent. Elke conceptvraag werd gevolgd door een bijbehorende isomorfe (gelijke) vraag. Deze studie toont aan dat leerlingen die feedback krijgen van de docent, of peerdiscussies voeren in combinatie met feedback van de docent, significant meer leerwinst behalen bij het beantwoorden van een isomorfe vraag in vergelijking met leerlingen die geen feedback krijgen. Deze leerlingen passen hun opgedane kennis toe op nieuwe isomorfe vragen. De meeste leerwinst wordt geboekt wanneer peerdiscussies gecombineerd worden met docentfeedback. De bevindingen tonen aan dat leerlingen die zouden kunnen leren van peerdiscussies (omdat ze de conceptvraag eerst fout beantwoordden, maar dit corrigeerden na peerdiscussies) inderdaad leren van peerdiscussies. Deze leerlingen beantwoorden verhoudingsgewijs meer isomorfe vragen correct dan leerlingen die geen feedback krijgen of niet met hun naaste medeleerlingen discussiëren, maar die de conceptvraag wel correct beantwoorden. Leerlingen kunnen kennis delen en (re)construeren in peerdiscussies terwijl ze voortbouwen op hun eigen concepten of op de ideeën van medeleerlingen, wat leidt tot een extra vergroting van begrip en omzetting van opgedane kennis naar nieuwe vervolg isomorfe vragen.

*Hoofdstuk 4* beschrijft een experimentele studie die de effecten onderzoekt van zowel docentfeedback, als van peerdiscussies in combinatie met docentfeedback op metacognitieve vaardigheden en motivatie. Gedurende een periode van tien weken beantwoordden zeshonderddrieëndertig natuurkundeleerlingen van zes scholen in het voortgezet onderwijs elke week vier gekoppelde sets van meerkeuzevragen met behulp van het webgebaseerde SRS *Socratic*. Leerlingen in de individuele groep en in de coöperatieve groep beantwoordden de meerkeuzevragen individueel. Echter, leerlingen in de coöperatieve groep bespraken hun antwoorden met hun naaste medeleerlingen voordat het antwoord werd uitgelegd door de

docent. Leerlingen in de controlegroep hoorden alleen het juiste antwoord, bespraken hun antwoorden niet met hun medeleerlingen, en kregen geen feedback van de docent. Deze studie toont aan dat peerdiscussies in combinatie met feedback van de docent een significant positief effect hebben op zowel metacognitie als motivatie. Het effect wordt gedeeltelijk gemedieerd door motivatie. Een verklaring voor dit mediërende effect is dat peerdiscussies formatieve toetsen interactiever en interessanter maken. Leerlingen worden pro-actiever en zijn meer gefocust bij het beantwoorden van meerkeuzevragen, wat leidt tot een verhoogde motivatie en, indirect, tot meer metacognitieve vaardigheden. Verder verschillen de effecten tussen meisjes en jongens, omdat meisjes hun metacognitieve vaardigheden verhogen en meer gemotiveerd raken door docentfeedback (al dan niet in combinatie met peerdiscussies) dan jongens, terwijl jongens hun metacognitieve vaardigheden en motivatie alleen verhogen door peerdiscussies in combinatie met docentfeedback. Meisjes hebben een positievere houding ten opzichte van feedback van de docent en zijn zich meer bewust van de kwaliteit van deze feedback, terwijl jongens hun metacognitieve vaardigheden verbeteren wanneer ze aangemoedigd worden om hun gedachten te bespreken. Hoofdstuk 4 laat ook differentiële effecten zien tussen leerlingen met verschillende metacognitieve vaardigheden. Laag-metacognitieve leerlingen hebben meer baat bij peerdiscussies in combinatie met docentfeedback dan hoog-metacognitieve leerlingen.

Veel docenten gebruiken SRS in hun formatieve toetsen zonder zich te realiseren in welke mate deze toetsen het leren van leerlingen beïnvloeden. Omdat formatieve toetsen met SRS metacognitief bewustzijn beïnvloedt, geeft *Hoofdstuk 5* inzicht in relaties van factoren die metacognitie kunnen beïnvloeden. De studie beschrijft het *cue utilization framework* van Koriat (1997) en het *control model* van Nelson en Narens (1990), en vult dit aan met aspecten die betrekking hebben op formatieve toetsing met SRS in een nieuw ontwikkeld framework. Het onderzoek toont aan dat formatief toetsen meerdere *prompts* bevatten, welke activiteiten zijn die leerlingen stimuleren cognitieve of metacognitieve strategieën aan te spreken. De meest voorkomende *prompts* in formatieve toetsingen met SRS zijn *anoniem stemmen*, *histogram feedback*, *feedback van de docent*, *peerdiscussies* en *isomorfe vragen*. Deze *prompts* stimuleren de voorkennis van leerlingen (de ene *prompt* wat meer dan de andere *prompt*) en helpen hen om één of meer diagnostische *cues* te herkennen die bepalend zijn voor het latere leren en presteren. *Cues* zijn informatie die de aandacht van de leerlingen



## Samenvatting

verleggen naar de vereiste of benodigde kennis. Veel voorkomende *cues* in formatieve toetsingen zijn *begrip cues*, *toets eisen cues*, *motivatie om te leren cues* en *angst cues*. Het ontwikkelde framework stelt dat *prompts* tijdens een toetsing leiden tot diagnostische *cues*, die de nauwkeurigheid van het beoordelen van monitoring verbeteren. Formatieve toetsingen kunnen op verschillende manieren worden uitgevoerd; een docent kan ervoor kiezen om histogrammen te tonen of niet, feedback te geven of niet, peerdiscussies toe te staan of niet, isomorfe vragen te stellen of niet, et cetera. Het aantal *prompts* dat leerlingen krijgen en de variëteit hiervan hangt dus af van hoe een docent een formatieve toets organiseert. Deze studie concludeert dat meer *prompts* tijdens een formatieve toetsing leiden tot meer diagnostische *cues*, wat resulteert in zowel een verbetering van de nauwkeurigheid van het beoordelen van de monitoring als van een verbetering van de metacognitieve vaardigheden. Wanneer docenten *prompts* samen gebruiken, zijn ze mogelijk nog effectiever omdat ze *cues* versterken bij het beoordelen van de monitoring en de regulatie van het leren.

In *Hoofdstuk 6* worden de bevindingen en conclusies uit de vorige hoofdstukken samengebracht in vijf hoofdconclusies. De eerste hoofdconclusie is dat het geven van feedback in formatieve toetsen zowel het begrijpen van de leerstof op korte als op lange termijn verhoogt. Leerlingen die feedback krijgen van de docent op conceptvragen, al dan niet in combinatie met peerdiscussies, verhogen hun kennis van de leerstof op de korte termijn. De resultaten impliceren effectgroottes van  $d = 0.83$  voor docentfeedback en  $d = 1.13$  voor peerdiscussie gecombineerd met docentfeedback. Daarnaast krijgen leerlingen die dagelijks formatief getoetst worden en feedback krijgen op hun antwoorden op de lange termijn meer inzicht in de leerstof dan leerlingen die deze toetsing en feedback op vragen niet krijgen. Dit lange-termijn inzicht komt overeen met een effectgrootte van  $d = 0.34$ . De tweede hoofdconclusie is dat leerlingen leren van peerdiscussies en daardoor hun metacognitief bewustzijn verbeteren. Leerlingen leren van peerdiscussies en kopiëren niet simpelweg hetzelfde antwoord van hun meer vaardige medestudent. Peerdiscussies stellen leerlingen in staat om te oefenen met het uitleggen van concepten aan elkaar, en antwoorden of oplossingen te ontdekken die ze zelf misschien niet zouden vinden. Peerdiscussies gecombineerd met feedback van de docent hebben ook een positief effect op metacognitie in vergelijking met leerlingen die geen feedback krijgen. De derde hoofdconclusie is dat de invloed van feedback op metacognitie verschilt tussen jongens en meisjes en tussen

leerlingen met een verschillend niveau van metacognitieve vaardigheden. Ten eerste zijn er verschillende effecten tussen jongens en meisjes. Vergeleken met jongens nemen bij meisjes de metacognitieve vaardigheden toe via zowel feedback van de docent als via een combinatie van peerdiscussies en docentfeedback, terwijl bij jongens de metacognitieve vaardigheden alleen toenemen via peerdiscussies in combinatie met docentfeedback. Op de tweede plaats zijn er verschillende effecten tussen subgroepen van leerlingen met verschillende metacognitieve vaardigheden. Laag-metacognitieve leerlingen profiteren significant meer van peerdiscussies bovenop de feedback van de docent dan hoog-metacognitieve leerlingen. De vierde hoofdconclusie is dat meer begrip van de leerstof leidt tot meer metacognitief bewustzijn wanneer leerlingen discussiëren met hun medeleerlingen en feedback krijgen van hun docenten. Het begrijpen van de leerstof na alleen passief te luisteren naar de feedback van docenten is voldoende voor de korte termijn, maar is onvoldoende om het leren te monitoren en te reguleren op een manier die de metacognitieve vaardigheden vergroot. Het is in feite de combinatie van docentfeedback met peerdiscussies die het conceptueel begrijpen vergroot en het metacognitief bewustzijn verhoogt, aangezien peerdiscussies situaties creëren waarin leerlingen hun eigen denkprocessen tonen, uitleggen, bespreken en controleren. De vijfde hoofdconclusie is dat angstgevoelens van leerlingen voor natuurkunde en hun motivatie om te leren de effecten van formatieve toetsen mediëren. Gevoelens van angst voor natuurkunde en motivatie om te leren mediëren de effecten van *prompts* op respectievelijk leerprestaties en metacognitie. Deze mediaties betekenen dat *prompts* ook *cues* zoals angst en motivatie triggeren, die op hun beurt metacognitie en leerprestaties beïnvloeden.



## About the Author

François Molin, born in 1976 in Maastricht, is a physics teacher in Dutch secondary education. He studied physics and obtained a Master's degree in Physics in 1998, followed by a Master's degree in Physics Education at the Fontys University of Applied Sciences in 2000. Since that time he has been working for Onderwijsgemeenschap Venlo & Omstreken (OGVO). In addition to his teaching, he was coordinator of the Technasium at College Den Hulster for almost ten years and has been co-author of NOVA Physics (Malmberg) since 2012. In the meantime, he studied Evidence-Based Innovation in Teaching at Maastricht University and obtained his Master's degree in 2016. One year later, he started his PhD research on a part-time basis at Maastricht University. He presented his research at various international conferences and published his work in several international peer-reviewed journals.



## ROA Dissertation Series

1. Lex Borghans (1993), *Educational Choice and Labour Market Information*, Maastricht, Research Centre for Education and the Labour Market.
2. Frank Cörvers (1999), *The Impact of Human Capital on International Competitiveness and Trade Performance of Manufacturing Sectors*, Maastricht, Research Centre for Education and the Labour Market.
3. Ben Kriechel (2003), *Heterogeneity Among Displaced Workers*, Maastricht, Research Centre for Education and the Labour Market.
4. Arnaud Dupuy (2004), *Assignment and Substitution in the Labour Market*, Maastricht, Research Centre for Education and the Labour Market.
5. Wendy Smits (2005), *The Quality of Apprenticeship Training, Conflicting Interests of Firms and Apprentices*, Maastricht, Research Centre for Education and the Labour Market.
6. Judith Semeijn (2005), *Academic Competences and Labour Market Entry: Studies Among Dutch Graduates*, Maastricht, Research Centre for Education and the Labour Market.
7. Jasper van Loo (2005), *Training, Labor Market Outcomes and Self-Management*, Maastricht, Research Centre for Education and the Labour Market.
8. Christoph Meng (2005), *Discipline-Specific or Academic? Acquisition, Role and Value of Higher Education Competencies*, Maastricht, Research Centre for Education and the Labour Market.
9. Andreas Ammermüller (2007), *Institutional Effects in the Production of Education: Evidence from European Schooling Systems*, Maastricht, Research Centre for Education and the Labour Market.
10. Bart Golsteyn (2007), *The Ability to Invest in Human Capital*, Maastricht, Research Centre for Education and the Labour Market.
11. Raymond Montizaan (2010), *Pension Rights, human capital development and well-being*, Maastricht, Research Centre for Education and the Labour Market.

12. Annemarie Nelen (2012), *Part-Time Employment and Human Capital Development*, Maastricht, Research Centre for Education and the Labour Market.
13. Jan Sauermann (2013), *Human Capital, Incentives, and Performance Outcomes*, Maastricht, Research Centre for Education and the Labour Market
14. Harald Ulrich Pfeifer (2013), *Empirical Investigations of Costs and Benefits of Vocational Education and Training*, Maastricht, Research Centre for Education and the Labour Market.
15. Charlotte Büchner (2013), *Social Background, Educational Attainment and Labor Integration: An Exploration of Underlying Processes and Dynamics*, Maastricht, Research Centre for Education and the Labour Market.
16. Martin Humburg (2014), *Skills and the Employability of University Graduates*, Maastricht, Research Centre for Education and the Labour Market.
17. Jan Feld (2014), *Making the Invisible Visible, Essays on Overconfidence, Discrimination and Peer Effects*, Maastricht, Research Centre for Education and the Labour Market.
18. Olga Skriabikova (2014), *Preferences, Institutions, and Economic Outcomes: an Empirical Investigation*, Maastricht, Research Centre for Education and the Labour Market.
19. Gabriele Marconi (2015), *Higher Education in the National and Global Economy*, Maastricht, Research Centre for Education and the Labour Market.
20. Nicolas Salamanca Acosta (2015), *Economic Preferences and Financial Risk-Taking*, Maastricht, Research Centre for Education and the Labour Market.
21. Ahmed Elsayed Mohamed (2015), *Essays on Working Hours*, Maastricht, Research Centre for Education and the Labour Market.
22. Roxanne Amanda Korthals (2015), *Tracking Students in Secondary Education, Consequences for Student Performance and Inequality*, Maastricht, Research Centre for Education and the Labour Market.
23. Maria Zumbuehl (2015), *Economic Preferences and Attitudes: Origins, Behavioral Impact, Stability and Measurement*, Maastricht, Research Centre for Education and the Labour Market.

24. Anika Jansen (2016), *Firms' incentives to provide apprenticeships – Studies on expected short- and long-term benefits*, Maastricht, Research Centre for Education and the Labour Market.
25. Jos Maarten Arnold Frank Sanders (2016), *Sustaining the employability of the low skilled worker: Development, mobility and work redesign*, Maastricht, Research Centre for Education and the Labour Market.
26. Marion Collewet (2017), *Working hours: preferences, well-being and productivity*, Maastricht, Research Centre for Education and the Labour Market.
27. Tom Stolp (2018), *Sorting in the Labor Market: The Role of Risk Preference and Stress*, Maastricht, Research Centre for Education and the Labour Market.
28. Frauke Meyer (2019), *Individual motives for (re-)distribution*, Maastricht, Research Centre for Education and the Labour Market.
29. Maria Ferreira Sequeda (2019), *Human Capital Development at School and Work*, Maastricht, Research Centre for Education and the Labour Market.
30. Marie-Christine Martha Fregin (2019), *Skill Matching and Outcomes: New Cross-Country Evidence*, Maastricht, Research Centre for Education and the Labour Market.
31. Sanne Johanna Leontien van Wetten (2020), *Human capital and employee entrepreneurship: The role of skills, personality characteristics and the work context*, Maastricht, Research Centre for Education and the Labour Market.
32. Cécile Alice Jeanne Magnée (2020), *Playing the hand you're dealt, The effects of family structure on children's personality and the effects of educational policy on educational outcomes of migrant children*, Maastricht, Research Centre for Education and the Labour Market.
33. Merve Nezihe Özer (2020), *Essays on drivers and long-term impact of migration*, Maastricht, Research Centre for Education and the Labour Market.
34. Inge Ingeborg Henrica Maria Hooijen (2021), *Place attractiveness, A study of the determinants playing a role in residential settlement behaviour*, Maastricht, Research Centre for Education and the Labour Market.
35. Alexandra Marie Catherine de Gendre (2021), *Behavioral Barriers to Success in Education*, Maastricht, Research Centre for Education and the Labour Market.



36. Kim van Broekhoven (2021), *From creativity to innovation: Understanding and improving the evaluation and selection of ideas in educational settings*, Maastricht, Research Centre for Education and the Labour Market.
37. François Molin (2022), *Using digital formative assessments to improve learning in physics education*, Maastricht, Research Centre for Education and the Labour Market.

