Trends in Biology Higher Education

A journey from interactive videos to pre-labs to rubrics

Marjolein E. Haagsman

Trends in Biology Higher Education

A journey from interactive videos to pre-labs to rubrics

ISBN: 978-94-6416-708-5
Cover design by Marjolein Haagsman
Layout by Marjolein Haagsman

Printed by Ridderprint

Trends in Biology Higher Education

A journey from interactive videos to pre-labs to rubrics

Trends in Biologie Hoger Onderwijs

Een reis langs interactieve video's, pre-labs en rubrics

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de Universiteit Utrecht
op gezag van de rector magnificus, prof.dr. H.R.B.M. Kummeling,
ingevolge het besluit van het college voor promoties in het openbaar
te verdedigen op woensdag 13 oktober 2021 des middags te 12.15 uur

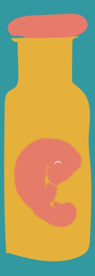
door

Marjolein Elisabeth Haagsman
geboren op 4 augustus 1987
te San Francisco, Verenigde Staten van Amerika

Promotor:	Prof. dr. J. Boonstra
Copromotoren:	Dr. M.C. Koster
	Dr. K. Scager

Table of Contents

1



Introduction

Introduction

1

The field of biology is ever changing. In the 17th and 18th century, biologists mainly performed descriptive research to classify species. In the 19th century, other disciplines such as medicine and physiology obtained a more prominent role, whereas molecular biology and systems biology are currently among the major disciplines in biology. One reason for the changes in the field of biology is that new research techniques became available to study other biological processes and structures. In addition, other biological questions become more relevant for society over time, so are the reduction of the carbon footprint and combatting the COVID-19 pandemic problems of our present time. The field of education is, similar to biology, subject to change. Education changes as the learning environment develops, other learning objectives become more relevant, new technical tools become available for learning and some learning mechanisms are better understood. Thus, just like in biology, there is a continuous need to further study, develop and improve our education.

Nonetheless, whereas biologists learn to think critically and base their arguments on thorough evidence, they rarely approach their own education as such (Handelsman et al. 2004). Teachers often do not have sufficient knowledge and time to reflect on their teaching or read educational studies to adapt their education. Thus, the gap between educational theory and practice remains. However, evidence-based education is becoming more popular, as shown by the increased interest in the Scholarship of Teaching and Learning (SoTL) (Tight 2018). The intention of SoTL research is to improve the teaching practice and assess whether the learning goals of a specific lesson activity are met (Hutchings & Shulman 1999). SoTL publications often focus on sharing best practices but not necessarily on improving our general understanding of learning. Even so, such SoTL studies can provide insights in the current issues of teaching and therefore also contribute to the direction of future education research on mechanisms of learning.

Next to SoTL, discipline based education research (DBER) is performed to study and improve learning in undergraduate science, technology and mathematics education (STEM) (Singer et al. 2011). The DBER is said to "investigate learning and teaching in a discipline using a range of methods with deep grounding in the discipline's priorities, worldview, knowledge, and practices. It is informed by and complementary to more general research on human learning and cognition" (National Research Council [NRC] 2012, p.9). Thus, DBER distinguishes itself from SoTL as it not only includes studies on reflective practice but also includes fundamental and theoretical studies to improve our general understanding of learning. In addition, DBER particularly focuses on students' skills and knowledge that are specifically required in science and on challenges that are specific for science education. A discipline within DBER is biology education research (BER), which aims to find out how students learn and become experts in the discipline of biology (Offerdahl et al. 2011). In the twentieth century, only 7% of papers and dissertations in science education were performed within this field, of which only half at college level and only a few by researchers with a biological background (DeHaan 2011). The field of BER is thus still relatively small, but the amount of papers and journals on biology education is rapidly growing (Dirks 2011; Lo et al. 2019). BER studies are important as they can contribute to the specific needs and issues in biology education.

In this thesis we describe different studies in the field of biology education research. The research questions relate to relevant topics and recent developments within this field, including the use of questions within educational videos, the employment of pre-labs and the use of rubrics for assessment. Although the topics are diverse, they all serve to improve the integration between practice and theory on learning.

Educational videos and pop-up questions

Videos are increasingly used in higher education for instructions and explanations, although the first videos for educational purposes were already designed in the 1950's, just after television was invented (Kay 2012; McIsaac & Gunawardena 2001). With the introduction of the first distance teaching universities in the 1970s, teachers in the United States started to develop video material they could share with their remote students with the aid of telecommunication delivery systems, but these delivery systems were costly. Hence, the use of video in education became more popular when wide distribution of video was facilitated by its digitalization in the late 1980s (Laaser & Toloza 2017). The digitalization of videos also made it possible to integrate text, images and sound. At this time, several universities established their own studio with specialized staff to record and edit videos for educational purposes. However, these videos were often only recorded for large educational projects because of their high costs. Some studios had to shut down when funding for multimedia learning stopped, resulting in a cutback on the use of video for education (Laaser & Toloza 2017). However, this rapidly changed from 2005 onwards when high speed bandwidth was invented and YouTube was founded, which made it very easy to share and view video material (Kay 2012). In addition, many devices acquired camera functions which made it possible for teachers, although perhaps of less quality, to easily record videos themselves. These advancements of video technology have led to the origin of massive open online courses (MOOCs) in 2008 and contributed to the increased use of the teaching method called the flipped classroom (Kaplan & Haenlein 2016).

The main idea of this flipped classroom strategy is that traditional in-class activities are now performed outside class and vice versa (Bishop & Verleger 2013). The traditional course-set up in higher education is that students learn new lesson content from lectures in class and then apply this information with assignments at home. Students in the flipped classroom, however, study the lesson content at home and apply this information with assignments in-class. The main idea of this approach is that students can learn the lesson content at their own pace at home and that teachers are present when students apply this information and probably when most help is needed. Educational videos are well suited for the flipped classroom since students can watch these videos at their own pace and time when studying the lesson content at home (Griffin et al. 2009; Hill & Nelson 2011; Hsin & Cigas 2013; Stockwell et al. 2015).

There are many diverse types of educational videos, such as webinars, animations, interviews, demonstrations, screencasts, recorded lectures, e-lectures, voice-over presentations and documentaries. Especially voice-over presentations are frequently used in higher education, probably because teachers in this field are already familiar with delivering lectures with slides (Griffin et al. 2009). Voice-over presentations that are short and only discuss one specific topic are also referred to as knowledge clips. The main motivation of using such shorter knowledge clips on one minor topic is that they are thought to reduce cognitive load of information, foster engagement and promote long term retention of the video content (Guo et al. 2014; Ibrahim 2011; Slemmons et al. 2018).

Watching these short knowledge clips is on itself a rather passive activity, but many websites and apps were recently developed to help enrich videos with interactive tools (HapYak 2021; Hihaho 2021; H5P 2021; Panopto 2021; PlayPosit 2021; Scalable Learning 2020). Examples of such interactive tools are chat fora, search tools, note-taking tools, marking tools and pop-up questions. These questions appear or "pop-up" at certain moments within a video and immediately need to be answered by the students before continuing the rest of the video. There is a long history of studies showing that testing increases retention of the tested information, and it is to be expected that pop-up questions within videos similarly promote learning (Karpicke & Blunt 2011; Pastötter & Bäuml 2014; Roediger & Butler 2011; Szpunar et al. 2008). However, little is known if pop-up questions indeed promote learning from video and how (different types of) pop-up questions affect the process of learning.

Laboratory activities and pre-labs

A very distinctive activity of science education is to perform experiments in the laboratory. For centuries, such lab activities are believed to be essential for learning science (DeBoer 2019a; Hofstein & Lunetta 2004). One of the early advocates for learning science in the laboratory was Thomas Huxley (Huxley 1899). Thomas Huxley was a British biologist with a strong belief in the need for teaching and studying science. Huxley thought that students should learn science by seeing and experiencing nature. He used botany as an example and explained that students will better understand the theory on plant anatomy if they dissect flowers themselves. Thus, Huxley claimed that the laboratory was essential to understand the theories and definitions explained in science textbooks. This belief was highly supported by Charles William Eliot in the late 19th century. Eliot was concerned that students were mainly taught in science from memorizing lectures and books, just as if they were studying grammar or history (Eliot 1898). He argued that science should be taught to students by doing observations. To Eliot's relief, laboratory activities became more common for science education during his time as president of Harvard University. At the same time, the field of biology research shifted from merely descriptive research towards more empirical research wherein experiments were designed. This also affected the beliefs on how lab activities should be taught. Some science teachers argued that students should learn science without much guidance of the teacher (Armstrong, 1898). Students should set up their own research questions, design their own methods and draw their own conclusions from the observations they make. Others argued that teachers should rather guide students, especially those in secondary education, through the research process with the aid of questions, hints and materials (Smith & Hall 1902). In the 1920s, there was a common belief that students would only learn if they were motivated to learn. Hence, it was commonly believed that the lab activities should promote students' interest by relating its content to the everyday life of the student. It became popular to design lab activities wherein students had to find a solution to everyday life problems with their own designed research (DeBoer 2019b). The problem of such "inquiry based" activities was that they cost a lot of time and money. Thus, more teachers started to give practical demonstrations instead, especially after some small studies showed that these demonstrations were just as effective as having students do the practical work. Soon ideas arose when to apply either teacher demonstrations or practical work by students (National Society for the Study of Education 1932). Practical work by students was thought to be essential if the goal was to develop lab techniques, but teacher demonstrations were thought to be merely as sufficient when the goal was to understand scientific principles. These beliefs on whether or when to use either teacher demonstrations, inquiry based labs or guided instruction labs have been an ongoing debate in the decades that follow (Hofstein & Lunetta 2004). In the 1990s, inquiry-based lab activities were once again promoted, although lab activities wherein students simply follow a certain procedure where still popular (Gunstone & Champagne 1990). Johnstone discussed how students are often expected to perform many different tasks when performing such "cookbook labs". Students are expected to follow procedures, recall scientific principles, search for materials, make observations, and analyze data at the same time (Johnstone, 1997). Thus, students are at risk of being overloaded with information when performing cookbook labs. Johnstone suggested that teachers should design pre-laboratory activities that could shape students' expectations of upcoming lab-activities and hence reduce cognitive overload in the laboratory. Such pre-labs can basically be any kind of lesson activity, such as face-to-face-tutorials, computer simulations, assignments or demonstrations. The current advent of instructional videos, wherein experimental proceedings are visualized, and virtual labs, wherein students perform experiments in a virtual environment, have increased the interest for using such pre-labs to reduce cognitive overload in the laboratory (Makransky et al. 2016; Nadelson et al. 2015). Such pre-labs are especially promising for biology labs that are designed to improve understanding of biological concepts by linking theory with practice. There is thus a need in biology higher education to find out if pre-labs can indeed enhance the outcomes of this learning objective.

Writing assignments and rubrics

Students in higher education commonly finish their studies with a written thesis on a specific topic in their field. The assessment procedure of written theses is rather complex since there is not one best approach or single answer for writing a good thesis (Sadler 2009). In addition, examiners often need to provide only one single grade but consider many different criteria such as significance, accuracy, scientific merit, use of primary literature and overall writing quality (Timmerman et al. 2011). In 1961, Diederich, French and Carlton emphasized the complexity of assessing such written assignments on writing ability. They let 53 examiners assess 300 theses and showed that 94% of these theses were assessed with at least seven different grades out of the nine grades that examiners could provide. They divided the examiners in five different groups based on their comments and assessments. The first group mainly commented and assessed on the relevance and the quality of the work, whereas the other groups mainly assessed on either the soundness, creative writing, mechanical errors and writing style. With these groups in mind, a framework of five criteria (ideas, form, favor, mechanics and wording) was designed to guide examiners in their assessment on writing ability. In the following years, the difference in assessments between examiners became a hot topic of discussion. More frameworks were designed to guide examiners in their assessment, all with the aim to increase the agreement between the assessment of different examiners. These frameworks, now commonly known as rubrics, generally include a list of assessment criteria with descriptions of when these criteria are considered as poor, excellent or somewhere in between. Examiners can then score every criterion separately before they decide upon the final overall grade. Indeed, some studies soon revealed how the use of these rubrics resulted in a higher level of agreement upon the final overall grade (Britton et al. 1966; Godshalk et al. 1966). Rubrics became truly popular in the 1990s, when it was commonly believed that assessments should not only measure students' abilities but also promote their own learning (Huot 1996; Wiggins 1993). Rubrics were thought to be a perfect instrument to provide this feedback since students could use them to evaluate and improve their own product (Dawson 2017). This explains why rubrics rapidly became a common instrument for assessment in higher education. To the current day, rubrics are still frequently used for assessments of assignments and many studies have measured how new rubrics are perceived by students and whether they affect the consistency of assessments (Reddy & Andrade 2010). Studies on the examiners' use of rubrics are however still rare, even though they can provide new insights on the quality of assessment.

Scope of this thesis

In this thesis I describe my research on the learning effects of pop-up questions, pre-labs and rubrics. The studies all start from a research question in the field of biology higher education. The obtained results help to advice teachers on how to improve their education and increase our current understanding of how learning and assessment occurs.

Learning from interactive voice-over presentations (Chapter 2 and 3)

In Chapter 2 and 3, studies are described on how pop-up questions improve learning from video in a flipped classroom setting. In Chapter 2, experiments on the learning effects of such pop-up questions are discussed. The learning mechanisms of pop-up questions are further explored in Chapter 3, were pop-up questions are compared that either preview upcoming video content (pre-questions) or review previous video content (post-questions). This chapter describes studies that examine both the learning mechanisms and learning effects of pre- and post-questions.

Understanding theory from laboratory activities (Chapter 4)

In Chapter 4, studies are presented that investigate if a pre-lab improves students' understanding of the theory and its relation to an experiment in the lab. A pre-lab computer module is developed that includes information on the background theory of the lab activity and contains questions on the purpose of each protocol step and on the interpretation of data. The studies in this chapter examine students' theoretical knowledge at the start of a laboratory activity, students' focus during the lab activity, and students' understanding of the experiment after the lab activity.

Assessing bachelor theses with rubrics (Chapter 5)

Bachelor theses are generally assessed by many different examiners with a large variety in expertise. These examiners have the difficult task to consider many different criteria when assessing theses with only one single grade. The grading procedure is often assisted with a rubric, but examiners are free in how to use the rubric for determining the final overall thesis grade. In Chapter 5, studies are described that explore which criteria correlate most with the overall assessment of the bachelor theses.

1

2

Pop-up questions within educational videos:

effects on students' learning

Marjolein E. Haagsman[1], Karin Scager[2], Johannes Boonstra[1], Margot C. Koster[1]

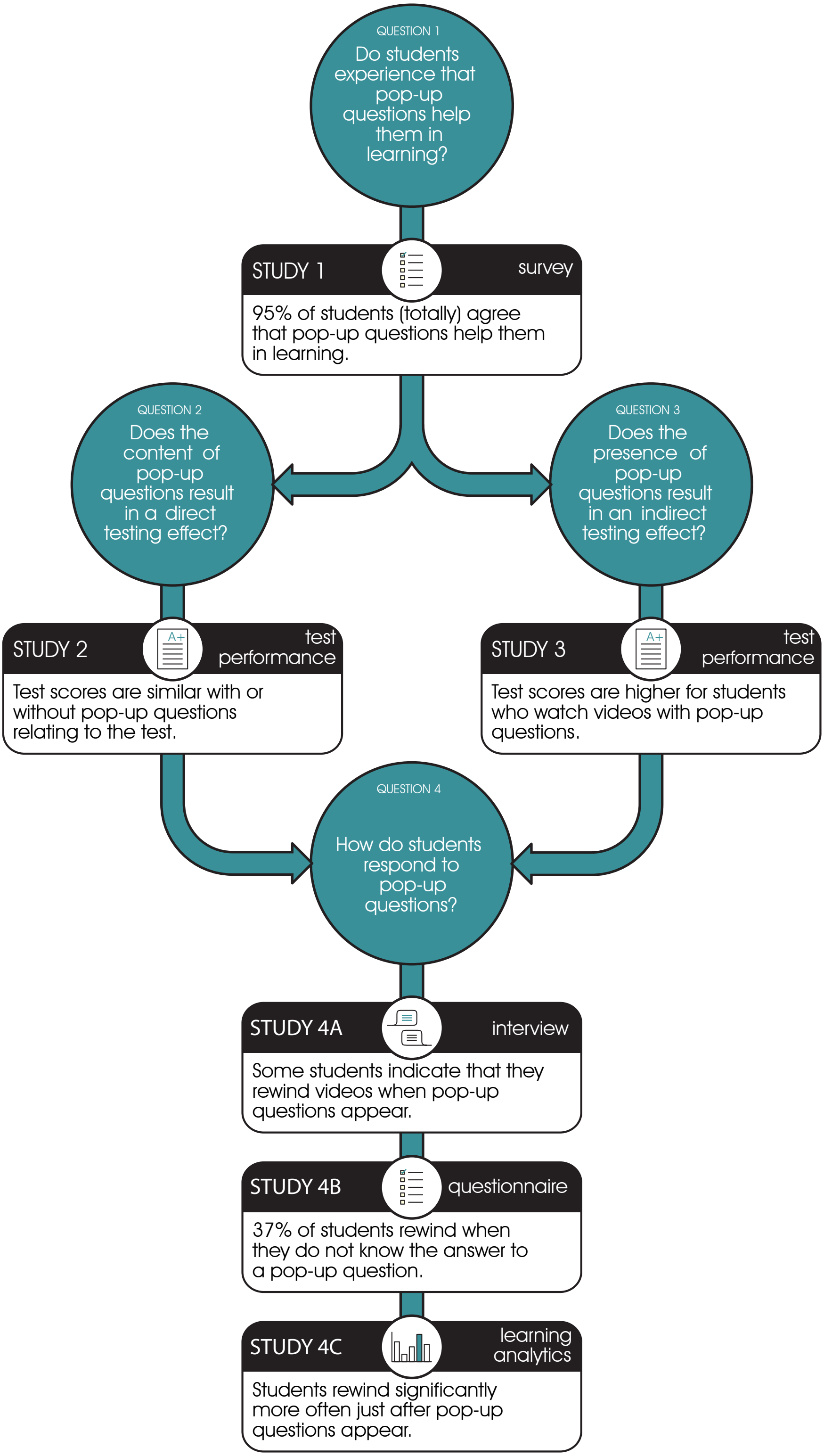
based on

Journal of Science Education and Technology 2020, 29, 713-724

1Department of Biology, Faculty of Science, Utrecht University

2 Department of Education, Faculty of Social and Behavioural Sciences, Utrecht University

2


QUESTION 1
Do students experience that pop-up questions help them in learning?
STUDY 1
survey
95% of students (totally) agree that pop-up questions help them in learning.
QUESTION 2
Does the content of pop-up questions result in a direct testing effect?
QUESTION 3
Does the presence of pop-up questions result in an indirect testing effect?
STUDY 2
test performance
Test scores are similar with or without pop-up questions relating to the test.
STUDY 3
test performance
Test scores are higher for students who watch videos with pop-up questions.
QUESTION 4
How do students respond to pop-up questions?
STUDY 4A
interview
Some students indicate that they rewind videos when pop-up questions appear.
STUDY 4B
questionnaire
37% of students rewind when they do not know the answer to a pop-up question.
STUDY 4C
learning analytics
Students rewind significantly more often just after pop-up questions appear.

Summary

Educational videos are increasingly used to let students prepare lesson material at home prior to in-class activities in flipped classrooms. The main challenge of this teaching strategy is to stimulate students to watch these videos attentively before going to class. This chapter describes the use of questions that pop-up within relatively long educational videos of sixteen minutes on average and designed to enhance students' engagement and understanding when preparing for in-class activities. The effects of such pop-up questions on students' learning performance were studied within a flipped course in molecular biology. Students had access to videos with or without a variable set of pop-up questions. The experimental group with pop-up questions showed significantly higher test results compared to the group without pop-up questions. Interestingly, students that answered pop-up questions on certain concepts did not score better on items testing these specific concepts than the control group. These results suggest that merely the presence of pop-up questions enhances students' learning. Additional data from interviews, surveys and learning analytics suggest that pop-up questions influence viewing behavior, likely by promoting engagement. It is concluded that pop-up questions stimulate learning when studying videos outside class through an indirect testing effect.

Introduction

Educational videos are regularly used to study information at home in flipped classroom education (Bishop and Verleger, 2013). The main idea of this flipped classroom model is that traditional class activities are shifted or "flipped" to activities outside class and vice versa. Thus, students study the lesson content outside class, which is often done with the aid of educational videos. Afterwards, students use the knowledge on a higher cognitive level during in-class activities. The main aim of this set-up is that teachers are present when students apply the information and probably when most help is needed. Accordingly, flipped lessons have shown to improve students' test performance within Science, Technology, Engineering and Mathematics (STEM) higher education (Baepler et al. 2014; Gross et al. 2015; Lax et al. 2016; Barral et al. 2018).

Although these results of flipped lessons on learning performance are promising, some challenges remain (Lo and Hew 2017). Herreid and Schiller (2013) described two major challenges for flipped classroom education experienced by STEM teachers. The first, rather practical challenge is that teachers find it hard to obtain or design proper videos suitable for studying the lesson content at home. The second challenge is that some students are not prepared well enough for the in-class activities. This last challenge is a fundamental issue of the flipped classroom model since student preparation is a prerequisite for in-depth in-class activities. The current chapter aims to investigate whether students' learning outside the classroom can be improved through video design by using questions that pop-up within educational videos.

Some suggestions for effective video design for learning have been made from the perspective of cognitive theory of multimedia learning (Mayer 2002). This theory suggests that media in learning should prompt cognitive processing of the relevant information without overloading the processing capacity of students. This cognitive process is mainly thought to be promoted by two conditions in video design: segmentation and signaling (Ibrahim 2011). In this model, segmentation is defined as the division of videos into smaller segments while signaling encompasses visual and audial signs that increase students' focus on the most relevant information. As such, students that watch videos with clear signals and small fragments are expected to have less attention to irrelevant information and will better remember the relevant information (Ibrahim 2011).

One highly promising tool to increase the effectivity of educational videos is the introduction of questions that pop-up within the video (Szpunar et al. 2013, 2014; Cummins et al. 2016; Lavigne and Risko 2018). These questions can either be in the form of single interspaced pop-up questions or as so-called interpolated tests with multiple questions. In either case, questions are included at certain intervals within an educational video and are expected to promote active engagement and thereby learning (Kumar 2010; Brame 2016). When used at regular intervals during the video, these pop-up questions hold both segmenting and signaling functions as adviced by Ibrahim (2011). The high interest in pop-up questions as a learning tool is reflected in the number of companies that provide tools to enrich videos with integrated questions (HapYak; Hihaho; H5P; Panopto; PlayPosit; Scalable Learning).

One of the early studies on videos with questions revealed that psychology students achieved higher performance on video-related test questions when watching the video with guiding questions on a separate sheet of paper (Lawson et al. 2006). Furthermore, several studies show that interpolated tests within or between videos do improve students' final test performance (Szpunar et al. 2013; Vural 2013; Lavigne and Risko 2018). The interpolated tests even appeared to be more effective on test performance than extra study time (Szpunar et al., 2013). In contrast, in a study of Wieling and Hofman (2010), interpolated tests did not affect students' final test performance within a course on European Law.

The positive effect of interpolated tests on final test performance observed in some studies could relate to the retrieval or testing effect, which is the finding that taking or practicing tests in general improves retention of information (Glover 1989). This testing effect has been supported by many studies and can be explained by both direct and indirect effects (e.g. Karpicke and Blunt 2011;

Pastötter and Bäuml 2014; Roediger and Butler 2011; Szpunar, McDermott, and Roediger 2008). The direct effect of testing occurs when testing enhances retention on a specific tested topic (Jacoby et al. 2010). One explanation for the direct testing effect is that students need to retrieve and process specific information when doing tests (Roediger and Karpicke 2006a). A second possible mechanism for the direct testing effect of pop-up questions in particular is that these questions operate as a signaling tool by recapitulating and testing the most relevant video content.

Besides enhancing retention on the specific tested topic, testing has also shown to enhance retention on subsequent non-tested lesson material (Chan et al. 2006; Szpunar et al. 2008). This indirect testing effect implies that also factors other than re-exposure and retrieval contribute to improved learning performance from tests. Recently suggested mechanisms for indirect testing effects of pop-up questions are an increase in note-taking (Lawson et al., 2006; Szpunar et al. 2013) and spending more time on the online learning material (Vural 2013). Moreover, students have reported to be more focused after each video fragment when they were tested during the videos (Szpunar et al. 2013), suggesting that the questions function as a segmentation tool. Summarizing, previous studies on videos with integrated questions suggest that they might promote learning both directly and indirectly by helping students to focus on the tested and most relevant information, process the tested information more elaborately, retain attention and stay actively involved. The number of studies on pop-up questions is however rather scarce and results are inconclusive.

More insight into the effect of pop-up questions on learning may help teachers to design effective videos. Such insights can be essential for flipped classroom education since the success of this model depends on the preparation by the students. In this study, educational videos on molecular biology are used of about sixteen minutes on average. These rather long educational videos are segmented with questions that pop-up about once per five or six minutes. The aim of this chapter is to examine whether pop-up questions enhance students' learning outside class within a flipped course in molecular biology. This study specifically aims to address the following three questions:

1. *Do students experience that pop-up questions help them in learning the video content?*
2. *Does the content of pop-up questions result in a direct testing effect?*
3. *Does the presence of pop-up questions result in an indirect testing effect?*

Based on the results from these studies we performed an additional explorative study to address the question:

4. *How do students use pop-up questions?*

The studies were performed in an authentic setting, meaning that the students watched the videos at home, while data were obtained from tests, surveys, interviews and learning analytics.

Methods

For this study a multimethod evaluation design was used. Qualitative and quantitative methods were employed sequentially, using the results of one method to design the next. First, students' perception of the effect of interactive videos on their learning was measured using an evaluation survey in 2015 **(Table 1)**. Second, in 2016, we studied direct testing effects of specific pop-up questions on the understanding of corresponding concepts. As a third step in 2017, we studied indirect testing effects from overall test performances. Based on results from these studies, we performed an extra study using focus group interviews to explore how students make use of pop-up questions. This outcome similarly resulted in a next study to explore students' viewing behavior with the aid of questionnaires. The result of this examination ultimately led to a final study to measure effects of pop-up questions on students' viewing behavior from learning analytics data.

▼**Table 1** Descriptive statistics for studies on effects of pop-up questions on students' learning.

Study	Research Question	Instrument	Number of participants N (%)	Year of Data Collection
Study 1	Do students experience that pop-up questions help them in learning the video content?	Survey	168 (69%)	2015
Study 2	Does the content of pop-up questions result in a direct testing effect?	Test performance	253 (94%)	2016
Study 3	Does the presence of pop-up questions result in an indirect testing effect?	Test performance	170 (57%)	2017
Study 4A	How do students use pop-up questions?	Interview	14 (8%)	2017
Study 4B	What do students do when pop-up questions appear?	Questionnaire	118 (69%)	2017
Study 4C	Do students rewind the video more often after pop-up questions appear?	Learning analytics	244 (82%)	2017

Participants

The participants in this study were freshman students on the Molecular Biology course (Department of Biology, Utrecht University). The first study on students' perception included 168 participants (69% response rate). The second study on the direct testing effect included 253 participants (94% response rate). The third study on the indirect testing effect was conducted with 170 (57% response rate) participants. The resulting extra studies on viewing behavior from interviews, questionaires and learning analytics included respectively 14 (8% response rate), 118 (69% response rate) and 244 participants (82% response rate). Note that the total number of students differs per experiment as the experiments were conducted over a period of three years. The participants within the comparative studies were randomly divided among experimental groups. Descriptive statistics on these groups can be found in **Table 2**.

▼**Table 2** Descriptive statistics for comparative studies on effects of pop-up questions on students' learning.

Year	Comparative study	Instrument	Group	Response rate	%male	%female	Average second exam grade (SD)[a]
2016	Study 2	Direct testing effect	Group A	149 (95%)	42%	58%	5.6 (1.50)
			Group B	104 (94%)	38%	62%	5.9 (1.44)
2017	Study 3	Indirect testing effect	Control group	83 (56%)	33%	67%	6.2 (1.52)
			Experimental group	87 (59%)	39%	61%	6.5 (1.46)
	Study 4C[b]	Rewinding behavior	Control group	121 (82%)	47%	53%	6.0 (1.62)
			Experimental group	123 (83%)	43%	57%	6.3 (1.48)

[a] All comparative studies were performed within the first part of the course. The presented grade is the average exam grade for the second part of the course. Students receive exam grades within a range of 1.0 (lowest) to 10.0 (highest).
[b]The demographic information and exam grades for study 4C are given for the total number of students instead of the respondents as data on rewinding behavior are anonymous.

Course design

The study was performed within the freshman course Molecular Biology, taught at Utrecht University in the Netherlands. The course is given in Dutch and is compulsory for all students participating in the undergraduate program of biology. The course content was based on the Text Book Biology, A Global Approach; Chapters 2–13 and Chapters 16–20 (10th and 11th International Edition) (Campbell et al. 2015, 2017). Research on video use was only performed within the first five weeks of the course. During this time, students were provided with four to eight videos per week. Students could view the videos voluntarily at home, at their own pace and in their own time. Additionally, understanding of the video content was tested weekly in online tests and then applied in group assignments. All tests, assignments and answers to pop-up questions were discussed weekly with the teacher during obligatory in-class activities in groups of approximately 40 students.

Video design

The educational videos were recorded by the teacher of the course. The videos were recorded as screencasts of slides with audio and lasted, on average, 16 minutes. The topics discussed within the videos were atoms and molecules, chemistry of water, carbon chemistry, biological macromolecules and lipids, energy, cell structure and function, cell membranes, cell signaling, cell cycle, cell respiration and photosynthesis. The videos were linked to the online video platform ScalableLearning (Scalable Learning 2020) to include pop-up questions within pre-made videos.

Pop-up questions design

Educational videos were enriched with pop-up questions reviewing the previously explained concepts. In 2015, the questions within the videos popped up once per eight minutes, on average. In the consecutive year, the main teacher added extra questions to the video as students reported that they would like to have more of them. Extra questions were added up to one question per five or six minutes, on average, depending on the experimental group. The questions were designed at the conceptual knowledge level of Bloom's taxonomy (Bloom et al. 1956). The videos paused when pop-up questions appeared within the video and automated feedback was provided after answering the question. Students that viewed the video clip for the first time could only continue the video after the pop-up question was answered correctly. However, students that rewatched the video clip could continue watching the video at any time by pressing the play button. The number of attempts and correct answers is provided to the teacher for the group as a whole but not per specific student. During the video clip, students also had the opportunity to use additional interactive tools. These tools included making digital notes, asking questions to the teacher and/or fellow students and pressing the "I am confused" button to label video fragments they did not understand. Students could also rewind, fast-forward, pause and change the speed of the video.

Test design

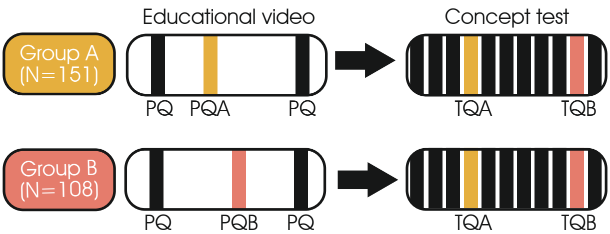
Online tests were designed to practice the concepts explained in the videos. The tests were, similarly to the pop-up questions, designed at the comprehension level of Bloom's taxonomy (Bloom et al.,1956). Furthermore, these tests were also used to measure students' learning performance for study 2 and 3 discussed further on. Students were asked to do eight tests of approximately 20 questions each. The tests were performed digitally at home and the deadline for these tests was one day before the corresponding in-class activities. The average score of the eight tests accounted for 5% of the final course grade.

Study 1 - Exploring students' experience on the effect of pop-up questions on their learning

In 2015, videos within the Molecular Biology course were embedded in ScalableLearning for the first time. Students' general perception of (interactive) videos was explored using a survey at the end of the course. Students responded to statements on a 5-point Likert Scale, ranging from strongly disagree (1) to strongly agree (5). The survey contained 14 questions on how students used these interactive tools within the video platform and whether these tools affected their learning. Only the following three statements concerning pop-up questions and the learning effect of educational videos are considered within this chapter: *The video clips helped me in learning*; *Answering the pop-up questions helped me in learning*; and *I would like to have fewer questions within a video clip*.

Study 2 - Measuring the direct testing effect

Students were randomly divided into two groups (A and B) at the start of the course. Each of these groups used a different course environment for watching the educational videos. The videos within the course environments were the same, but 14 extra pop-up questions were inserted for alternating groups **(Figure 1)**.

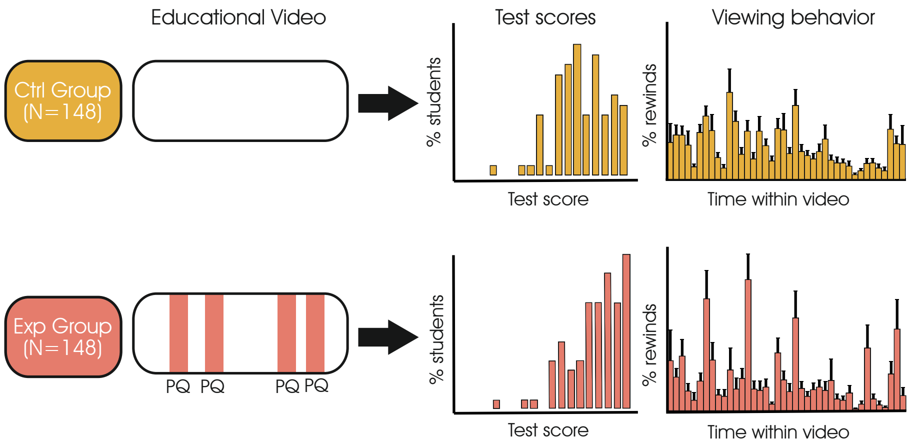


▲**Figure 1** Schematic design for measuring understanding of concepts (study 2). In this experiment, students were divided into group A and B. Both groups viewed the same videos with different pop-up questions (PQ). The corresponding test questions (TQ) in the test are marked similarly.

The pop-up questions were based on the course learning goals and developed on the level of comprehension. Corresponding test questions were designed for each of those questions and were incorporated into the tests covering the entire study content. The test questions were not identical to the pop-up questions but tested the same concept at the same comprehension level. For example, one pop-up question was: "*Which of the following amino acids does not contain asymmetric carbon atoms?*" Whereas, the corresponding test question contained a structural formula with the question: "*Which of the carbon atoms within the structural formula below is asymmetric?*". Each individual score per specific test question was obtained for comparison. Only test scores were analyzed for students that attempted to answer the corresponding pop-up question. Students who did the digital test before fully watching the corresponding videos were excluded from analysis in order to obtain a solid measurement of the effect of video preparation on test performance. The remaining students did the tests with a median time interval of 1 day 6 hours and 58 minutes after watching the videos.

Study 3 – Measuring the indirect testing effect

In the subsequent year, students were randomly divided between an experimental and a control group. Both groups watched two educational videos on cell signaling, with durations of, respectively, 20:10 minutes and 19:33 minutes. Four pop-up questions were designed for both educational videos. However, in this experiment only one experimental group received these pop-up questions whereas the other control group received no pop-up questions at all **(Figure 2)**.



▲**Figure 2** Schematic design for measuring students' overall test performance (study 3) and viewing behavior (study 4C). In this experiment, students were divided into an experimental and a control group. Both groups viewed the same videos on cell signaling. The experimental group received pop-up questions (PQ) within the videos, whereas the control group did not receive any pop-up questions. Students' test performances on the overall video content were compared between the experimental and control group. The groups were also compared on the number of rewinds and fast-forwards per student.

Students' general conceptual understanding of these two videos was tested with a corresponding test on cell signaling. The scores of this test were compared between the control and experimental group and corrected for other test scores obtained prior to the experiment. Again, students were excluded from the analysis when they did the digital test before fully watching the corresponding videos.

Study 4A - Exploring students' use of pop-up questions via focus groups

After the Molecular Biology course in 2016, two groups of six and eight students participated in a semi-structured focus group interview on their use of pop-up questions. Students were asked to describe their actions when questions appeared within the video.

Study 4B - Exploring students' use of pop-up questions via questionnaires

The results of the focus group interviews were used to design a questionnaire on the use of videos and pop-up questions. The questionnaire contained closed questions on students' use of videos, and the multiple-choice answers to these questions were derived from student discussions during the focus group interviews. Only one question concerning students' behavior when not knowing the answer to a pop-up question is used in this chapter (N=118). Other questions concerning students' general use of video were considered to be irrelevant for the current chapter.

Study 4C - Measuring students' rewinding behavior

Viewing behavior was analyzed for the same video clips on cell signaling **(Figure 3)**. The specific data used for this study were the number of rewinds per student. Only rewinds of more than one second were used for analysis. The percentage of rewinds per student was determined for every timeframe of 30 seconds in the video. The total use of rewind and fast-forward buttons within the video was also determined for the experimental and control group and compared with a control clip. This control clip was a video clip of 18:07 minutes on cell structure and function which contained five pop-up questions that were identical for both the control and experimental groups. The raw data on rewinds and fast-forwards were provided personally by the development team of the video platform ScalableLearning.

Statistical analysis

For the first step of the analyses we performed descriptive statistics. The answers to the test questions within study 2 were scored as either correct or incorrect, and Pearson's chi-square analysis was performed to compare these results for groups with or without corresponding pop-up questions. For study 3, a one-way analysis of covariance (ANCOVA) was conducted to compare test scores of the experimental and control group. The test scores were controlled for the second exam grade of the course. An independent t-test was used for study 4C to compare the mean percentages of rewinds within the 30 seconds after pop-up questions between the control group and experimental group. The average number of rewinds and fast-forwards throughout the entire videos was not normally distributed and compared with a Mann-Whitney test. Individual rewinds and fast-forwards greater than three times the interquartile range of each experimental group were considered as outliers and removed from this analysis. All statistical analyses were performed using IBM SPSS Statistics Version 24.

Results

Student perception on the effect of interactive video on their learning

The present study started with a general student evaluation of the interactive video platform. A few questions within this survey examined whether students believed that educational videos and pop-up questions helped them in learning **(Table 3)**.

Table 3 shows that 97% (totally) agreed that video clips in general helped them in studying the learning content. In addition, 91% of the students (totally) agreed that pop-up questions, specifically, helped them in studying. This positive attitude towards pop-up questions was confirmed by the finding that 79% of the students (totally) disagreed with decreasing the number of questions within the video.

2

▼**Table 3** Descriptive statistics from an evaluation survey on interactive video for items on video clips and pop-up questions (study 1, N=168)

Statement	Likert Scale Response (%)					*N*	Mean	S.D.
	1 Totally disagree	2 Disagree	3 Neutral	4 Agree	5 Totally agree			
The video clips helped me in learning	0 (0%)	0 (0%)	4 (2%)	51 (31%)	108 (66%)	163	4.64	0.53
Answering the pop-up questions helped me in learning	0 (0%)	3 (2%)	11 (7%)	95 (61%)	48 (31%)	157	4.20	0.65
I would like to have fewer pop-up questions within a video clip	38 (24%)	87 (55%)	22 (14%)	9 (6%)	1 (1%)	157	2.03	0.82

The numbers in each category represent the numbers of students answering in that category.
The mean (M) and standard deviation (S.D) presented are the mean and standard deviation values derived from the Likert Scale ranging from 1 (totally disagree) to 5 (totally agree). The sum of the percentages is not equal to 100% due to rounding errors.

Direct testing effect

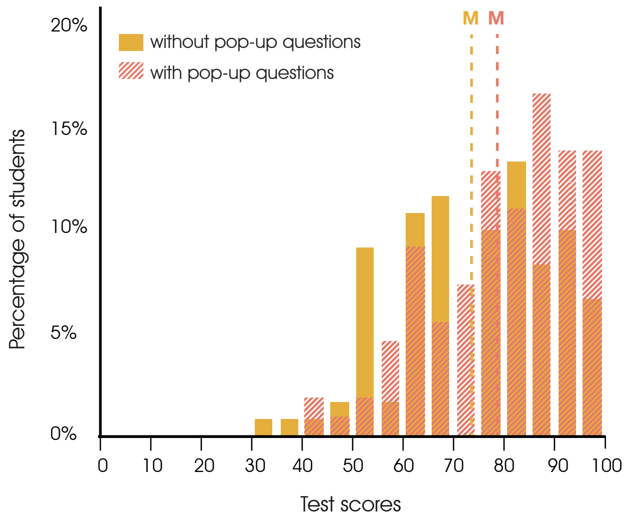
Tests were performed prior to the in-class activities to investigate whether a pop-up question on a specific concept helped students to understand that specific concept. An experimental set-up was designed in which two groups watched the same video clips with different pop-up questions on different concepts. Afterwards, both groups did a test on the video content. The test scores on the individual items are shown in **Table 4.** Surprisingly, the percentage of correctly answered test questions was not significantly different between students that did (72%) or did not (69%) receive corresponding pop-up questions (χ^2(1, N=2901) =2.52, p=0.11). Students with corresponding pop-up questions only performed significantly better on one question (item 5), which was the only question that was nearly identical to the pop-up question (χ^2 (1, N=181) =15.10, p<0.001).

▼**Table 4** Test scores of students with or without access to corresponding pop-up questions (Study 2, N=253).

		Without pop-up question		With pop-up question				
Experimental group	Test Item	N[a]	Correct	N[a]	Correct	χ^2	*df*	*p*
Group A	2	102	44%	97	50%	0.575	1	0.448
	4	85	68%	123	64%	0.359	1	0.549
	6	94	51%	128	54%	0.404	1	0.525
	7	78	72%	103	65%	0.928	1	0.335
	10	81	75%	107	80%	0.694	1	0.405
	12	90	86%	114	86%	0.007	1	0.934
	14	70	81%	89	85%	0.450	1	0.502
Group B	1	148	95%	103	95%	0.038	1	0.846
	3	139	94%	103	98%	2.172	1	0.196[b]
	5 [c]	103	45%	79	73%	15.098	1	<0.001***
	8	124	90%	89	96%	2.003	1	0.157
	9	121	30%	84	29%	0.033	1	0.855
	11	89	67%	70	66%	0.051	1	0.821
	13	109	57%	85	57%	0.003	1	0.954

Indirect testing effect

In the previous experiment, group A and B were both required to answer pop-up questions although on different concepts. A follow-up experiment was performed to examine whether merely the presence of pop-up questions might affect student performance on the full video content. In this experiment, students watched a video on cell signaling either with or without pop-up questions. Afterwards, students were tested on the entire video content and their test scores compared with an ANCOVA. Interestingly, there was a significant effect of the presence of pop-up questions on these overall test scores after controlling for their exam grade **(Figure 3)**. Students who watched videos with pop-up questions scored significantly better on the test (M_{adj}=79%, SE=1.17) than students who watched the same video without pop-up questions (M_{adj}=75%, SE=1.11); F(1,218)=7.68, p=0.006 **(Supplementary S1)**. Students with pop-up questions particularly scored more often above 85% when pop-up questions were present **(Figure 3).**



▲**Figure 3** Percentage of students with distinct test scores for groups that watched a video clip on cell signaling with or without pop-up questions (study 3). The possible test scores ranged from 0 to 100. The dashed lines represent the mean adjusted test scores for both the control group (N=83) and the experimental group (N=87)

Students' use of pop-up questions

The previous two experiments suggest that pop-up questions do not improve test performance on the specific tested concept, but that merely the presence of pop-up questions affects test performance on the video content as a whole. These results motivated us to perform a set of extra studies and explore possible causes of indirect testing effects. Two semi-structured focus group interviews were performed to investigate how students use pop-up questions. First, students were asked how, where and when they were watching the video. Some students watched the video when commuting in the train or bus but most students watched them at home. Some students explained that they watched the video in one go whereas others said they used their phone or computer at the same time:

STUDENT 8: *It's not like I am continuously on my phone or something. No. Otherwise I could not follow the story.*

STUDENT 9: *I sometimes turn that thing (educational video) on and then I do something else on my computer in the meantime. (…) I always try to do a bit of multitasking.*

Students were then asked to describe their first actions when a question pops up and what they did when their answer to a pop-up question was incorrect. Some students commented that they simply tried the next answer, as they explained:

STUDENT 4: *Most often when I receive a question, I just give the answer that I think is right and then I just try the next. It's not like I look back for those things.*

A few students specifically clarified that they guessed because they wanted to continue listening to the video lecture:

STUDENT 9: Yes, *I do this as quickly as possible because you want to continue the rest of the thing. So you quickly think about it…*

Other students explained that they rewind the video when they do not know the answer to a pop-up question, although one explains that he/she only does this when preparing for final exams and not when preparing for in-class activities:

STUDENT 10: *For me it is an indicator of understanding the previous fragment. Usually, when I answer incorrectly, I rewind part of the video.*

Similar results were found from the subsequent questionnaire, showing that 47% of all students indicated that they guessed the answer until they found the correct one **(Table 5)**. About 37% of the students indicated rewinding the video first when not knowing the answer to the question. The remaining students claimed to search for the answer on the Internet or in the textbook.

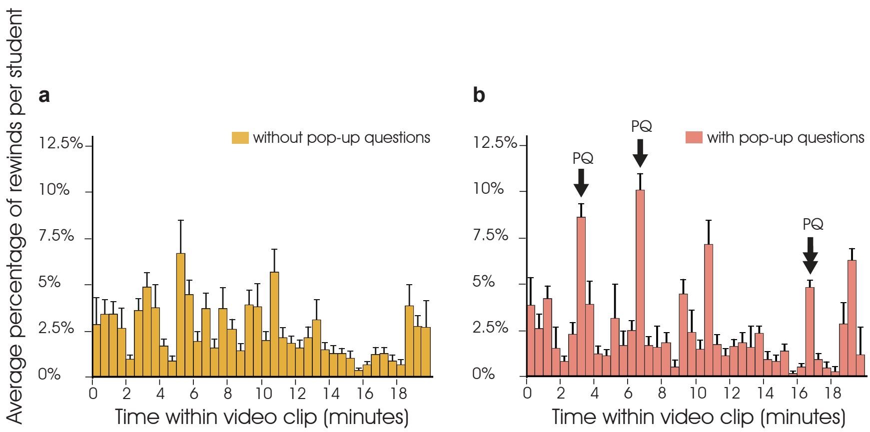
▼**Table 5** Students' responses to the statement "What do you do first when you do not know the answer to a pop-up question?" (Study 4B, N=118)

		Likert Scale Response (%)			
Item	Statement	Guess answer	Rewind video	Search on Internet	Search in textbook
5	What do you do first when you do not know the answer to a pop-up question?	47%	37%	9%	6%

The sum of the percentages is not equal to 100% due to rounding errors

In order to get more insight into the influence of pop-up questions on students' rewinding behavior, we used learning analytics data. We determined the use of the rewind buttons for both the experimental group with pop-up questions and the control group without pop-up questions. The effect of pop-up questions on rewinding behavior was analyzed from the relative number of rewinds through the course of a video clip **(Figure 4)**. These results reveal that students rewind relatively more often within the 30 seconds after pop-up questions occur ($M = 0.22$, $SD = 0.23$) as compared to the same time points in the control video without questions ($M=0.10$, $SD=0.12$); $t(168)=-4.535$, $p<0.001$. Similar results were found for a comparable video clip **(Supplementary S2)**.

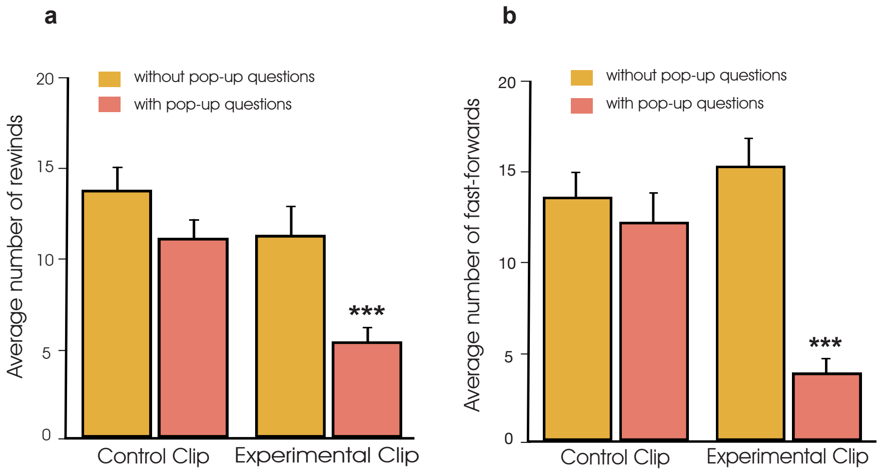
We also explored the effect of pop-up questions on the general use of both rewind and fast-forward buttons throughout the entire video. Interestingly, students in the experimental group (with pop-up questions) rewound significantly less (Mdn=3) compared to the control group (Mdn=8), U= 3946.50, p<0.001 **(Figure 5A)**. In addition, students also fast-forwarded significantly less when pop-up questions were present (Mdn=0) as compared to when no pop-up questions appeared (Mdn=5), U=3524.00, p<0.001 **(Figure 5B)**. Similar results were found for a comparable video clip **(Supplementary S3)**.



▲**Figure 4** Average percentage of rewinds per student within the time course of a video clip with (a) and without (b) pop-up questions (study 4C). The percentages ofrewinds are shown for every 30 s ofthe video for both the control group (N = 117) and the experimental group (N = 120). The arrows illustrate time points of the pop-up questions shown to the experimental group at 3m22s, 6m31s, 16m39s, and 16m57s. Error bars represent standard errors of the means.

No significant difference in the average number of rewinds was found for a control clip similar for both groups. Thus, students rewind and fast-forward less often throughout the video clip as a whole, although they do rewind more often just after pop-up questions appear.

2



▲**Figure 5** Average number of rewinds (a) and fast-forwards (b) during videos with and without pop-up questions (study 4C). The experimental video clip contained four pop-up questions for the experimental group (N=120) and no pop-up questions for the control group (N= 117). The control videoclip contained four pop-up questions for both the control group (N=132) and the experimental group (N=127). Error bars represent standard errors of the means.

Discussion

The study demonstrates that pop-up questions within educational videos improve students' test performance on the overall video content. Accordingly, students agreed that pop-up questions within educational videos helped them to study at home and were positive about including more pop-up questions within the videos. However, pop-up questions on a particular concept within the video did not improve test performance on that specific concept. Thus, our pop-up questions did not result in a direct effect, but rather in an indirect effect on students' test performance.

It is surprising that we did not find a direct testing effect for pop-up questions, since such an effect has been reported by several previous studies (Butler 2010; Glover 1989; Karpicke and Roediger 2008; Szpunar et al 2008). One explanation is that studies on the direct testing effect mainly addressed the memorization of vocabulary lists in which students are tested at the level of remembering of Bloom's taxonomy (Bloom et al. 1956). In the present study, however, students were tested at the level of comprehension. Interestingly, one of the test questions accidently appeared to be designed for the level of remembering since it was nearly identical to the corresponding pop-up question itself. This pop-up question was also the only question that resulted in a significantly higher score of the corresponding test question. This finding suggests, although speculative, that pop-up questions and answers were remembered but simply not improved students' comprehension of that specific concept.

This study was performed in an authentic setting, meaning that we could only slightly control how students watch the educational videos and how they answer the tests. Therefore, one limitation of this study is that we could not control whether students used any help when performing tests at home. One other limitation of this authentic setting is that data acquisition occurred over multiple years, leading to subtle differences in the course set-up between experiments. Nonetheless, conclusions were only drawn by comparing results of groups within one cohort. Students were randomly divided over these groups and group results on test performances were corrected for differences in their knowledge.

Future studies are required to determine whether differences in cognitive levels of pop-up questions affect learning differently. It is however unlikely that the use of questions at the level of evaluation will improve the conceptual understanding, since Cummins et al. (2016) reported low study engagement for such pop-up questions. These results were confirmed by the students in our study, who claimed that they would not benefit from more difficult questions, as this would only stimulate them to guess and click through all of the possible answers until correct. Nonetheless, we recommend future studies to investigate different parameters of pop-up questions that might result in direct testing effects such as the level and the frequency of pop-up questions or whether pop-up questions either review or preview the video content.

The lack of a direct effect on concept understanding may also be partly explained by the following explorative studies showing that nearly half of the students claimed to guess the answer to a pop-up question and thus did not review the learning content. Some of the students did search for the right answer, either by rewinding or studying other sources, although the effect of this more dedicated approach did not show in the test results. The learning analytics data confirmed that students rewound the video more often just after a pop-up question. A similar effect has been reported before in a study on text instead of videos (Rouet et al. 2001). Rouet et al. provided online texts to students and recorded their scrolling behavior. Interestingly, these students appeared to reread previous information more often when in-text questions were present. The authors of this study propose that text reviewing promotes and guides a deeper level of text comprehension. However, our study did not investigate the effects of the different approaches of students towards pop-up questions. It would be interesting to determine whether students that review the content, do have an increase in conceptual understanding when analyzed separately.

Although we do not report a direct testing effect for pop-up questions, we do show that merely the presence of pop-up questions promoted student performance in the test as a whole. Possible mechanisms for such indirect testing effects have been proposed from previous studies on interpolated tests between video fragments (Lawson et al. 2006; Szpunar et al. 2013; Vural 2013). For example, students have reported to show less mind-wandering when these interpolated tests were present (Szpunar et al. 2013). Although we did not specifically examine mind-wandering, our learning analytics data do reveal that students rewind and fast-forward less over the course of an entire video when pop-up questions are present. We hypothesize that this decrease in zapping back and forth through a video might actually be a result of a higher focus of attention, and this might be particularly true when using relatively long educational videos such as the videos in this study. Other previously reported mechanisms are more note-taking and spending more time on the learning material when video fragments are interpolated with tests (Lawson et alw 2006; Vural 2013). Just the presence of pop-up questions could hence increase students' attention to the video.
In conclusion, our results suggest that teachers can manipulate students' attention and (re-)viewing behavior by inserting pop-up questions within educational videos. Hence, pop-up questions can improve students' learning

when watching videos at home. This finding is of particular interest for teachers in a flipped classroom setting who design videos as a preparation for in-class activities.

Informed consent and statement of human rights

All procedures performed in this research, involving human participants, were in accordance with the ethical standards of the institutional and/or national research committee (Review Board of Social Sciences at Utrecht University, IRB approval number FETC180-962). An informed consent was obtained from all individual participants or anonymous data collection was used.

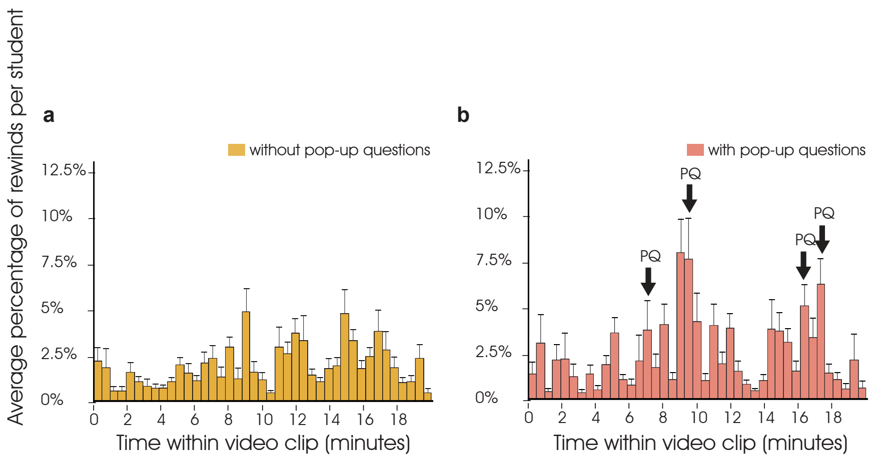
Supplemental tables and figures

▼**Supplementary S1** ANCOVA results and descriptive statistics for experimental test scores of groups that watched a video clip on cell signaling either with (experimental group) or without (control group) pop-up questions. The experimental test contained questions on cell signaling and were scored from 0% to 100%. The experimental test scores were adjusted for the second exam grades of the course.

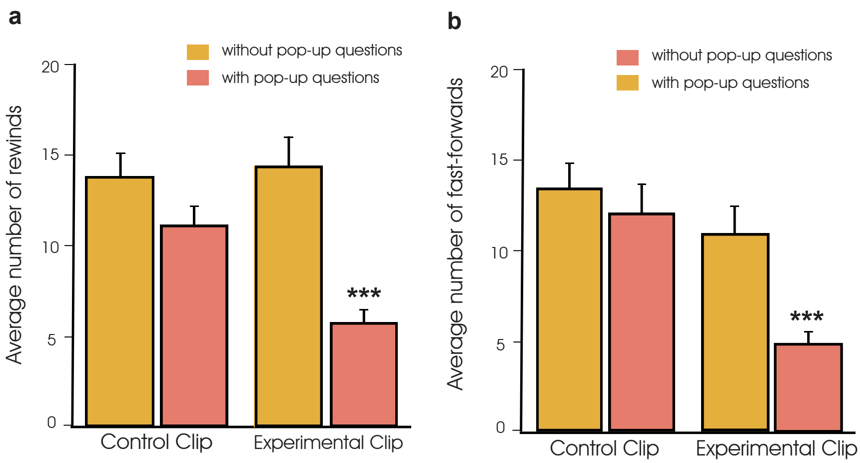
Group	Observed Mean	Adjusted Mean	S.E.	N	
Experimental Group	80%	79%	1.17	105	
Control Group	74%	75%	1.11	116	
Source	**SS**	**df**	**MS**	**F**	**p**
Second Exam Grade	14106	1	14106	98.24	<0.001***
Control Group vs Experimental Group	1103	1	1103	7.68	0.006
Error	31302	218	144		

Note. R2 = 0.34, Adj. R2 = 0.33. The adjustments are based on second exam grade = 6.3. Homogeneity of regression was tested and not significant: F = 0.58, p=0.362 The second exam grade regression coefficient = 5.38***.

***p<0.001



▲**Supplementary S2** Average percentage of rewinds per student within the time course of an educational video without (a) and with (b) pop-up questions (study 4C). The percentages of rewinds are shown for every 30 seconds of the video for both the control group (N=121) and the experimental group (N=123). The arrows illustrate the time points of the pop-up questions at 7m22s, 9m53s, 16m57s and 17m42s. Error bars represent standard errors of the means.



▲**Supplementary S3** Average number of rewinds (a) and fast-forwards (b) during videos with and without pop-up questions (study 4C). The experimental video clip contained four pop-up questions for the experimental group (N=123) and no pop-up questions for the control group (N=121). The control video clip contained four pop-up questions for both the control group (N=132) and the experimental group (N=127). Error bars represent standard errors of the means.

3

To preview or to review:

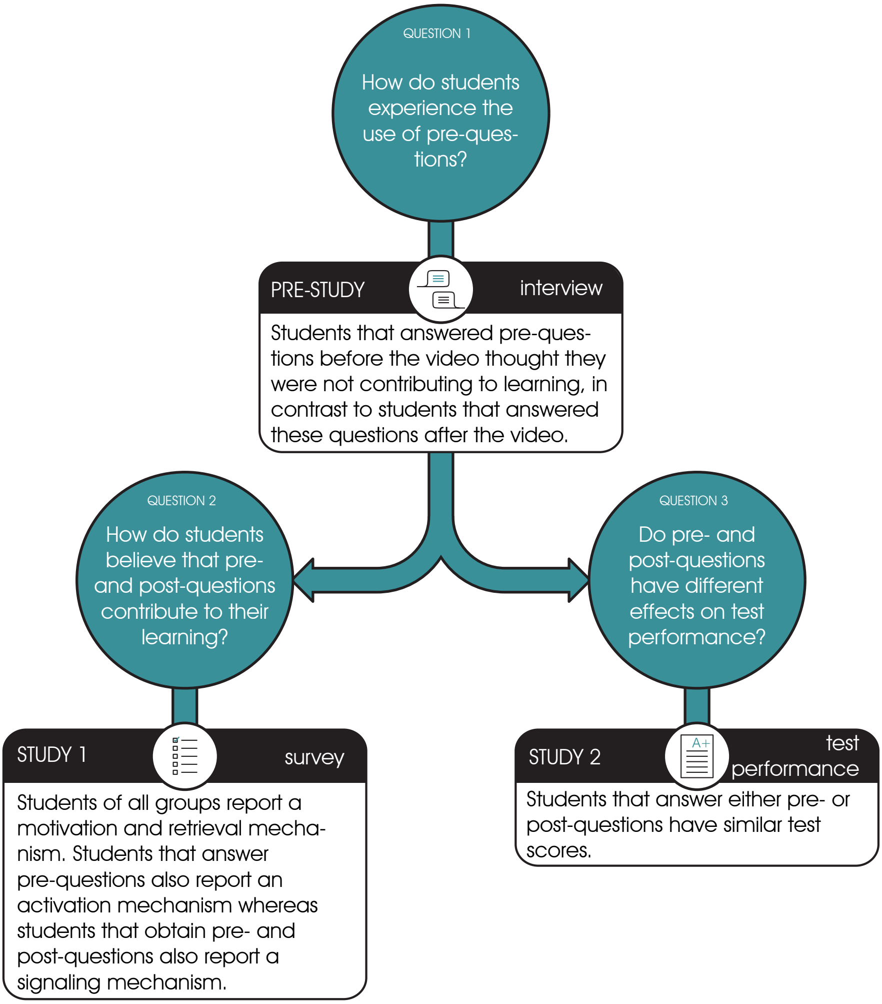
do pre-questions and post-questions direct students' learning from knowledge clips?

Marjolein E. Haagsman[1], Laurens A. Meeteren[1], Margot C. Koster[1], Johannes Boonstra[1], Karin Scager[2]

Submitted

1Department of Biology, Faculty of Science, Utrecht University

2 Department of Education, Faculty of Social and Behavioural Sciences, Utrecht University


QUESTION 1
How do students experience the use of pre-questions?
PRE-STUDY
interview
Students that answered pre-questions before the video thought they were not contributing to learning, in contrast to students that answered these questions after the video.
QUESTION 2
How do students believe that pre- and post-questions contribute to their learning?
QUESTION 3
Do pre- and post-questions have different effects on test performance?
STUDY 1
survey
Students of all groups report a motivation and retrieval mechanism. Students that answer pre-questions also report an activation mechanism whereas students that obtain pre- and post-questions also report a signaling mechanism.
STUDY 2
A+
test performance
Students that answer either pre- or post-questions have similar test scores.

Summary

In this chapter we investigate the use of questions inserted within educational videos designed to promote active participation. These "pop-up questions" can either be designed as pre-questions that preview upcoming video content or as post-questions that review previous video content. In this study we explored if different types of pop-up questions have different effects on learning and how students believe that these different questions promote their learning. The study was performed in a flipped undergraduate course in molecular biology. One group answered pre-questions, one group answered post-questions and one group read pre-questions and answered them later on as post-questions. Survey responses reveal that students from all groups believe that the pop-up questions improved their retrieval of the information and increased their focus on the video. Students that answered post-questions almost solely report these two learning mechanisms. However, the other groups also perceive different additional learning mechanisms for pop-up questions. Students that answered pre-questions explained in the survey that these questions helped them to activate their prior knowledge, although they were not so positive about the usefulness of this type of questions when interviewed. Students that read pre-questions and answered them later on as post-questions responded in the interview that the pop-up questions were useful and specifically explained in the survey that these questions signalled the most relevant parts of the video. Nonetheless, no differences are shown in test performances for students that answered either pre-questions, post-questions or both. These results suggest that, even though pre-questions and post-questions might promote different learning mechanisms, they have no different effects on the retention and understanding of the video content.

Introduction

Instructional videos are increasingly used to motivate, inform and instruct students in higher education (Kay 2012). Additional tools have been developed to make videos more interactive with the aim to stimulate active processing of information and increase their effectivity on learning (Derry et al. 2014). One of these tools is introducing questions that students are required to answer while watching the video. We refer to these questions as pop-up questions since they suddenly "pop-up" at specific time points. Several studies demonstrated that students gain higher test results on video content when videos were interpolated with these pop-up questions (Lavigne and Risko 2018; Haagsman et al. 2020; Szpunar et al. 2013; Vural 2013).

Previously, we showed that the presence of pop-up questions already improved students' test performance (Haagsman et al. 2020). Interestingly, the students that answered pop-up questions on certain concepts did not score better on items testing these specific concepts than the control group. These results lead to the conclusion that merely the presence of pop-up questions enhances students' learning, likely by promoting engagement. However, the type of questions used may also influence the effect on learning. The pop-up questions in our study were all designed to review information explained in the previous video fragment. Alternatively, questions may be used that preview the upcoming video content. We refer to these two types of questions as post-questions (questions that review the previous content) and pre-questions (questions that preview the upcoming content). Both pre-questions and post-questions in videos appear to promote retention of the video content (Carpenter and Toftness 2017; Toftness et al. 2018). Nonetheless, different learning mechanisms for these types of questions have been proposed as discussed below. We are therefore interested to determine if these types of questions have different effects on students' test performance and how students perceive the value of pre- and post-questions on learning from video.

Proposed learning mechanisms of post-questions

Post-questions are generally perceived by students as tools to measure their understanding of previous information (Karpicke 2009). However, post-questions not only appear to measure knowledge but also to increase retention of the tested information (Glover 1989). This learning effect is called the testing effect and has been demonstrated in many different settings; both for questions within texts and videos (Glover 1989; Karpicke and Blunt 2011; Roediger and Butler 2011; Roediger and Karpicke 2006a; Szpunar et al. 2008). Post-questions not only increased students' retention, but also appeared to improve the ability of students to apply the studied information (Butler 2010; Chan et al. 2006). Three mechanisms have been proposed to explain how post-questions might improve learning: the motivation mechanism, the reviewing mechanism and the retrieval mechanism.

The first proposed learning mechanism is the motivation mechanism, which suggests that post-questions result in a higher interest, focus and attention on the video content. The motivation mechanism is proposed from our previous study, wherein students zapped less often back and forth through a video when answering post-questions, possibly as a result from a higher focus of attention (Haagsman et al. 2020). This learning mechanism is also supported by a study from Szpunar et al. (2013), showing that students report less mind-wandering when answering post-questions. In addition, students that answer post-questions report more thoughts related to the online lecture than students that restudy the lecture content instead (Jing et al., 2016). Thus, these studies suggest that post-questions induce a motivation mechanism by promoting students' focus and interest on the video content.

The second proposed learning mechanism is the reviewing mechanism. Post-questions are expected to encourage students to review the lesson content, which ultimately results in additional study time. This mechanism is supported by a study from Vural (2013), showing that students indeed spent more time on the learning material when watching videos that include post-questions. Besides, our previous study showed that students rewinded videos more often just after a post-question, which suggests that these questions promote students to review part of the video (Haagsman et al. 2020). However, most studies that showed positive effects of post-questions have been performed in laboratory settings where rereading and rewinding could not occur, suggesting that the reviewing mechanism is not the only mechanism that explains how post-questions promote learning (Butler 2010; Chan et al. 2006; Glover 1989; Roediger and Karpicke 2006b; Szpunar et al. 2008).

The third proposed learning mechanism of post-questions is the retrieval mechanism. Students who answer post-questions need to recall or retrieve the previous studied information from their memory. This retrieval of information ultimately results in reconsolidation of the new information from the text or video (Antony et al. 2017; Lee et al. 2017). In other words, students who retrieve information from their memory are thought to re-encode that tested information which ultimately results in stronger memories and higher retention (Carpenter et al. 2008; Chan et al. 2006; Karpicke 2009; Karpicke and Blunt 2011; Karpicke and Roediger 2007). The retrieval mechanism is supported by a study from Pyc and Rawson (2009) on learning Swahili-English word pairs. Students practiced these word pairs until they were correctly retrieved for a certain number of times. The more often the students had to practice and recall a word pair from their memory, the better the word-pair was eventually remembered.

Proposed mechanisms of pre-questions

The previous suggested mechanisms on reviewing and retrieval are specific for questions that review previous information. Several studies have shown that pre-questions in texts can, just like post-questions, also improve retention of the discussed material (Bull and Dizney 1973; Kornell et al. 2009; Little and Bjork 2011; Richland et al. 2009). One recent study on pre-questions within video lectures similarly showed that students could recall more information from videos that include pre-questions than from videos without questions (Carpenter and Toftness 2017).

Pre-questions cannot lead to reviewing or retrieving previous given information, and it is thus expected that the learning effects of pre-questions result from different mechanisms in learning. Although pre-questions are studied less frequently than post-questions, a few mechanisms of the learning effects of pre-questions have been proposed. The first proposed learning mechanism is the activation mechanism. The activation mechanism states that pre-questions prime students' prior knowledge, resulting in a better integration of that knowledge with the upcoming information (Alba and Hasher 1983; Hannafin and Hughes 1986). The effects of prior knowledge on brain activity and learning are nicely shown in a study on participants that studied combinations of object words with visual patterns and pieces of fabric (Van Kesteren et al. 2010). Participants had to study both congruent (such as leather and jacket) and incongruent combinations (such as lace and umbrella). Congruent combinations were remembered significantly more often, and MRI results revealed that these combinations resulted in higher activity of the medial prefrontal cortex; a region of the brain known to be involved in retrieving information. Thus, these results are consistent with the notion that prior knowledge promotes learning of new knowledge. These studies suggest that pre-questions may promote learning by stimulating students to integrate the upcoming information with the knowledge they already have.

The second possible learning mechanism of pre-questions is the signaling mechanism. Pre-questions are thought to function as a signal for students by drawing their attention to the essential parts of the upcoming information (Hannafin and Hughes 1986; Rothkopf and Coke 1963). These effects of pre-questions are nicely presented in an early study from Mayer (1975). The participants in this study studied six pieces of texts on mathematics and were either tested on the definitions from the text, on their ability to perform calculations from the text, or on their ability to apply these calculations in a new type of problem. All participants then studied two additional texts, and all performed a test on these texts with all three types of questions. Interestingly, students performed significantly better on the definition questions if previously tested on definitions. Similarly, students performed significantly better on the calculation questions if previously tested on calculations. These results suggest that a pre-question can signal students to focus on a certain upcoming topic and thereby promote its learning. If teachers design pre-questions on topics, they find most relevant, pre-questions are expected to signal and promote learning of the relevant content.

The third proposed learning mechanism of pre-questions is the motivation mechanism, which is also expected to occur for post-questions as previously described. It is hypothesized that pre-questions specifically promote students' interest and attention on the upcoming information (Geller et al. 2017; Hannafin and Hughes 1986; Litman et al. 2005; Little and Bjork 2011). The motivation mechanism of pre-questions is supported by a study from Rotgans and Schmidt (2011) who showed that students report higher interest to their lessons when first presented with a related problem on the topic of that lesson. In addition, pre-questions appear to be more effective on learning if students report to be highly curious to know the answer (Bull and Dizney 1973). Thus, pre-questions are expected to increase curiosity, foster attention, and hence promote learning of the upcoming information.

In conclusion, pre-questions and post-questions both appear to induce retention of previous or upcoming information and several mechanisms have been proposed to explain these testing effects. Some of these mechanisms are only applicable to post-questions whereas others are only applicable to pre-questions. In the current study we aim to find out whether pre-questions and post-questions differ in their contribution to learning from video and how learners perceive their learning effect.

The participants of this study were freshman biology students that watched educational videos on molecular biology. The students were divided in different groups: a pre-question group and a post-question group. All groups answered identical questions within the video, but these questions were either asked before or after the corresponding theory in the video. An additional group of students obtained questions twice: once before the explanation and once after the explanation. The purpose of this group was to promote learning mechanisms of both pre-questions and post-questions. The following two questions were investigated:

1. *Does the type of question (pre-questions, post-questions or a combination of both) in educational videos affect students' learning performance?*
2. *How do students experience the contribution of pre-questions and post-questions on their learning from educational videos?*

Methods

Participants

The participants in this study were freshman students participating in the biology undergraduate program at Utrecht University (the Netherlands). A pre-study was performed in 2018-2019, in which students were randomly divided into two groups: a group that answered pre-questions (group 1) and a group that read pre-questions and answered post-questions (group 2). Four students of group 1 and six students of group 2 participated in an interview on their experiences of using pop-up questions within educational videos. In 2019-2020, students were randomly divided into three groups: 95 students that answered pre-questions (group 1), 95 students that read pre-questions and answered them later on as post-questions (group 2) and 102 students that answered post-questions (group 3). In both years, students were informed about the aims of the study and asked for their consent for analyzing and publishing anonymized data on their answers to test questions, survey questions and interview questions. Only participants that provided consent were included in the datasets of this study.

Course design

The study was performed within the first course of the undergraduate program Biology. This course is compulsory and discusses Chapters 2-13 and Chapters 16-20 of the textbook Biology, A Global Approach (11th International Edition) (Campbell et al. 2017). The Molecular Biology course is taught in a flipped classroom setting in which students watch videos at home that explain the theoretical information. Additionally, students' understanding of the video content is tested with individual online tests and then applied on campus with group assignments. Afterwards, the videos, tests and group assignments are discussed on campus with the teachers in separate groups of about 30 students. This sequence of events occurs once a week.

Video design

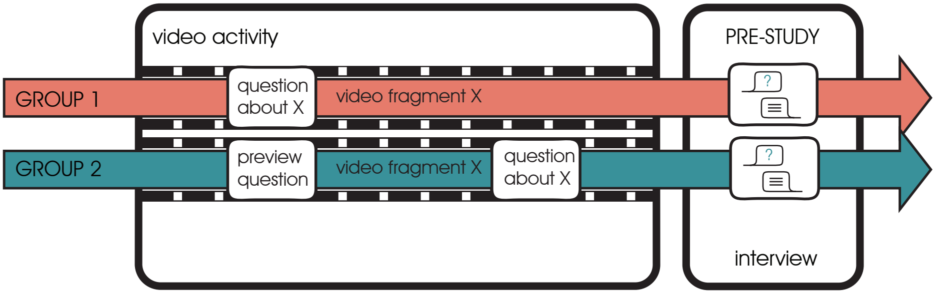
All videos were recorded as screencast of slides together with audio and/or video of the main teacher. The topics of the videos were designed to revive high school material on biological macromolecules, lipids and energy (pre-study) and on atoms and molecules, chemistry of water and carbon chemistry (study 1 and 2). The videos were linked to the online video platform Scalable Learning (Scalable Learning 2020). In this video platform, students could ask questions to their fellow students and/or teacher and press the "I am confused" button to show that they did not understand the topic discussed. Furthermore, teachers could insert pop-up questions within videos on this video platform.

Pop-up questions

The pop-up questions were inserted at specific moments in the video and videos paused when pop-up questions appeared. Students obtained automated feedback immediately after they answered a pop-up question. If their answer was incorrect, they were required to answer the question again before they could continue the video. If their answer was correct, students could continue watching the video. The pop-up questions were based on the learning objectives of the videos and designed at the comprehension level of understanding of Blooms taxonomy (Bloom et al. 1956).

Pre-study

The pre-study was performed during the third week of the course (year 2018-2019), when students already had two weeks of experience in watching videos and answering pop-up questions. During this week, students were required to watch 11 educational videos in their own time without supervision. These videos discussed topics on cell structure and function and cell membranes. The 11 videos were together 1 hour and 53 minutes long and included eight pop-up questions. The pop-up questions were the same for both group 1 (94 students) and group 2 (100 students) but were presented at different times within the videos. Students of group 1 received pre-questions prior to the explanation of the corresponding theory. Students of this group were required to immediately answer the question before continuing the video. Students of group 2 also received the pop-up question prior to the corresponding video content but only required to answer these questions after the corresponding video content **(Figure 1)**.

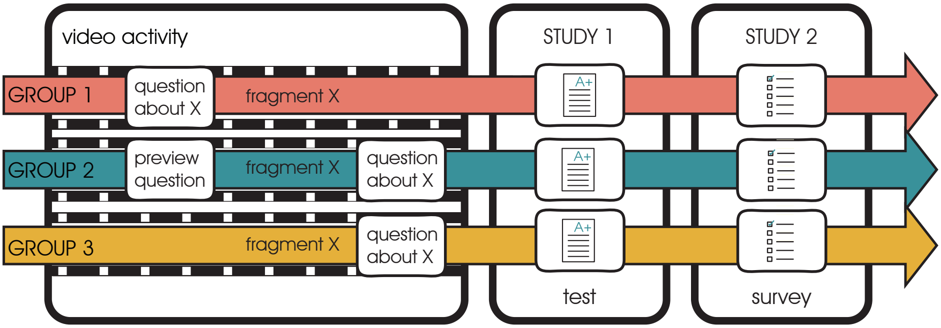


▲**Figure 1** Schematic presentation of the pre-study. Students of group 1 (pre-questions) and group 2 (pre- and postquestions) watched videos intersected with pop-up questions. For both groups, questions were posted before its corresponding video fragment (X). Students of group 1 immediately answered that question. Students of group 2 read the question before and answered the question after the video fragment. Afterwards, students were interviewed on their experiences on answering pop-up questions within videos.

Semi-structured interviews were performed on the same day that students applied the video content during group assignments. A set of guideline questions on students' experiences in answering pop-up questions were designed. The interviews were performed in two separate sessions: one session with four students of group 1 and one session with six students of group 2. The interviews were audio-recorded and answers to the questions were documented with notes and quotes.

Experimental set up

The pre-study was continued with an experiment on the first day of the course (year 2019-2020). During this day, students were required to watch five videos on campus under supervision of a teacher. The topics of these videos were atoms and molecules, chemistry of water and carbon chemistry. The five videos were together 1 hour and 1 minute long and included 12 pop-up questions. The pop-up questions were the same for group 1 (95 students), group 2 (95 students) and group 3 (102 students). However, for every group, pop-up questions were presented at different times within the videos. Students of group 1 were immediately required to answer pop-up questions prior to the corresponding video content **(Figure 2)**. Students of group 3 were immediately required to answer pop-up questions after the corresponding video content. Students of group 2 were presented with the pop-up question prior to the corresponding video content but only required to answer these questions after this video content. After watching the video, all groups were asked to do a test on the video content and respond to a questionnaire on the added value of pop-up questions. The pop-up questions, test and survey were all answered on campus under supervision of teachers. Additional help of others or textbooks was not allowed.



▲ **Figure 2** Schematic presentation of study 1 and 2. Students of group 1 (pre-questions), group 2 (pre- and post-questions) and group 3 (post-questions) all watched videos intersected with pop-up questions. Students of group 1 answered pop-up questions before its corresponding video fragment (X). Students of group 2 read the question before and answered the question after that video fragment. Students of group 3 read and answered the question after the corresponding video fragment. Afterwards, students of all groups performed a test on the video content and answered a survey on the benefits and shortcomings of pop-up questions in video.

Study 1 – Test performance

Immediately after watching the videos, students were asked to answer a set of test questions incorporated in the online video platform **(Figure 2)**. The test was, just like the pop-up questions, designed at the level of comprehension (Bloom et al. 1956). The test included 12 questions that tested the same concepts as the 12 pop-up questions within the video in similar ways. For example, one of the pop-up questions was as follows: '*An element (potassium) has an atomic number of 19 and a mass number of 39. How many protons, electrons and neutrons does this element have?*', whereas the corresponding test question read: *'An element contains 10 electrons, 10 protons and 12 neutrons. What is the atomic number and mass number of this element?'*. The test was answered by 287 students with informed consent; 94 students of the group 1, 92 students of group 2 and 101 students of the group 3. Test scores were compared with a one-way analysis of variance (ANOVA) using IBM SPSS Statistics Version 26.

Study 2 - Student survey

After the test, students were asked to answer a set of survey questions incorporated in the online video platform **(Figure 2)**. The survey included three closed questions and two open questions on the added value and quality of the pop-up questions. The closed questions included the following statements with a Likert rating scale from 1 (totally disagree) to 5 (totally agree): *'The pop-up questions helped me in learning'*, *'I would like to have fewer pop-up questions in the video'* and *'The pop-up questions were difficult'*. The first open question of the survey was: *'What did you like about the pop-up questions?'.* In total, 260 students answered to this question: 89 of the group 1, 82 students of group 2 and 89 students of group 3. In total, 58 student responses did not reflect any mechanisms of learning, but instead provided descriptive answers such as *"nice summary"* or *"not too difficult"* or *"questions were clear"*. These answers were identified by the first author and removed from analysis. The remaining answers were blindly coded with a deductive coding approach into one or more of the following pre-designed categories: reviewing mechanism, retrieval mechanism, activation mechanism, motivation mechanism and signaling mechanism **(Table 1)**. The deductive coding was performed in multiple steps. First, fifty responses were blindly assigned by the authors to one or more of these categories. Second, differences in annotation were discussed together to arrive at a final consensus. Next, the remaining responses were also blindly coded by the authors, after which the differences were again discussed to arrive at a final consensus. The frequency of answers per category were then compared between the three different groups using descriptive statistics in SPSS (statistics computer software).

The second open question of the survey was: *'What didn't you like about the pop-up questions?'*. In total, 197 students answered this question: 78 of group 1, 57 students of group 2 and 62 students of group 3. Answers to this question were blindly coded with an inductive coding approach (Thomas, 2006). First, categories were developed by the first author based on the student responses found. The answers were annotated into the following categories: less interruption, high number, other level, other content, less mistakes, better formulation, different timing, unnecessary preview, more feedback and technical issues **(see Supplemental Material 1)**. Second, student responses were blindly and independently coded by the first and fourth author to one or more of these categories. Third, differences in annotation were discussed together to arrive at a final consensus. The frequency of answers per category were then compared between the three different groups using descriptive statistics in SPSS (statistics computer software).

▼**Table 1** Categorization of students' responses on what they liked about pop-up questions.

Mechanism	Description	Example
Reviewing mechanism	students rewind the video or reread textbooks	"They encourage to review the video when answers are incorrect"
Retrieval mechanism	students retrieve the information through testing or rethinking about the video content	"It immediately makes you think about the content, immediately apply it"
Activation mechanism	students' prior knowledge is activated	"It helps you think about the topic before it is explained, which makes you participate more actively during the lesson"
Motivation mechanism	students are more focused and more interested in the video content	"They keep me focused"
Signaling mechanism	students shift their attention to a certain topic of the video	"The questions help me to select the most important content from the video."

Results

Interview responses

The first aim of the study was to explore the perception of the students on pop-up questions displayed before the corresponding video content. In this pre-study, all students were required to watch videos with such pop-up questions. One group of students immediately had to answer these questions (group 1) whereas the other group had to answer these questions only after the corresponding explanation in the video (group 2). The purpose of the set-up of group 2 was that students might improve learning through the proposed mechanisms of both pre-questions and post-questions. Students of group 1 and 2 were both interviewed separately on their experiences in answering these questions.

Students of group 1 thought that the pre-questions were not contributing to learning, since they felt that they did not have enough information to answer the question and would soon learn the answer from the video anyway, as one student explained: "*Now they are often posted before it and then they literally give the answer to these questions one second later. Then I think… Yes, I find this method of questioning quite useless.*" The students also explained that they did not have the urge to look for the answer to the pre-questions right away, since they first wanted to listen to the following explanation within the video: "*I would search for the answer only after I got information and don't know the answer, or if I want to delve into it. I think this is more useful; then you already watched a bit of video and understand what you get or don't get and can find out yourself what to do with it.*" The students did explain that pre-questions might be an advantage for easy questions, as one student clarifies: "*In principle it works for easy questions because then you see; I still understand the content. And then the upcoming content is also easier to understand. If you got it wrong you see; shit, that is wrong, and then you don't have any misconceptions in the content explained thereafter.*"

Students of group 2 explained that the pre-questions were pleasant as they immediately grabbed their attention and made them more actively involved: "*The first question puts you right in it. Then you immediately need to concentrate on the content*". Furthermore, students of this group explained that the pre-questions helped them in understanding the theory and directed their focus towards the video content related to the pop-up questions: "*I thought it was very pleasant because then you change your focus on what to think about, so this helps, at least for me, to get the goal of the clip clear*". Nevertheless, one student also noted that, especially after watching many or rather long videos, she did not always remember the pre-question when continuing the video.

Students explained that the obtained feedback from the corresponding post-questions made them aware of what they understood from the previous video content, as one student explained: "*The final question also serves as a warning: sometimes you think you understood but then you still got it wrong.*" Three out of six students also clarified that they kept record of the time points where each topic was explained in the video. They explained that this helped them to find and review a specific topic. The students described that, if they got the answer to this final question wrong, they would rewind the video to review that specific topic.

In other words, students that answer pre-questions generally experienced these questions as not so useful since they felt that they did not have enough knowledge to answer the question at that moment. In contrast, students that read pre-questions and answered them only after that video fragment were generally positive on the pop-up questions as they explained that they were more focused, more actively involved and better aware of what they did or did not understand from the video.

Survey responses

In a follow-up study, we conducted surveys to investigate whether pre- or post-questions result in different reported benefits of learning, suggesting different learning mechanisms to occur. Students had to watch videos and answer pop-up questions under supervision and without help of others. Moreover, students were now divided in three student groups. Students of group 1 were, similar to the pre-study, immediately required to answer pop-up questions before the corresponding video fragment. Students of group 2 were, similar to the pre-study, presented with pop-up questions before the corresponding video fragment but only required to answer them after this theory was explained. Students of group 3 only answered questions after the corresponding theory. After watching the videos, all groups were immediately asked to respond to a survey on what they believed was the added value of pop-up questions.

On a 5-point Likert scale (1-totally disagree to 5-totally agree), most students agreed that pop-up questions helped them in understanding the video content (M=3.98, SD=0.83) and disagreed that these pop-up questions were difficult (M=1.97, SD=0.53). Students were also asked to express the added value of pop-up questions in an open question. Overall, students most frequently reported a retrieval mechanism by explaining that the questions helped them to rethink about the video content or tested their understanding **(Table 2)**. In addition, students from all groups commonly reported a motivation mechanism by explaining that the questions increased their focus of attention on the video (group 1: 34%, group 2: 33%, group 3: 35%).

▼**Table 2** Number of students that report a learning mechanism of pop-up questions when asked about the added value of pop-up questions.

Mechanisms	Group 1 (n=65)	Group 2 (n=69)	Group 3 (n=68)
Reviewing mechanism	0 (0%)	1 (1%)	3 (4%)
Retrieval mechanism	26 (40%)	35 (51%)	39 (57%)
Activation mechanism	16 (25%)	1 (1%)	0 (0%)
Motivation mechanism	22 (34%)	23 (33%)	24 (35%)
Signaling mechanism	0 (0%)	11 (16%)	1 (1%)
Other	6 (9%)	5 (7%)	6 (9%)

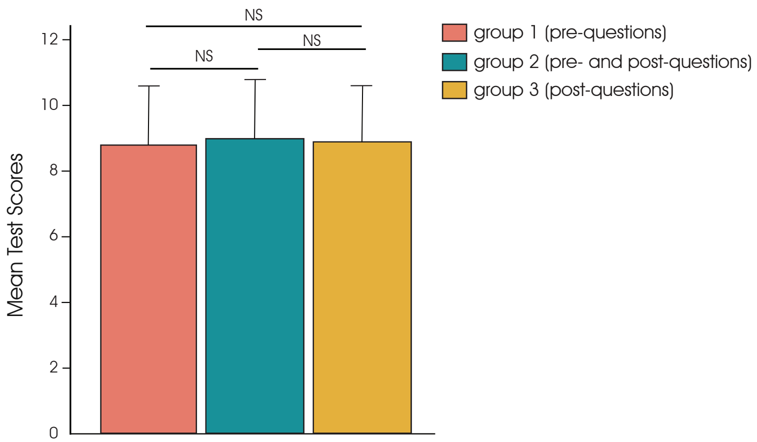
Student responses are shown for group 1 (pre-questions), group 2 (pre- and post-questions) and group 3 (post-questions). Note that some students reported more than one learning mechanism.

As expected, the retrieval mechanism is reported more frequently by students of group 2 (51%) and 3 (57%) than students of group 1 (40%) that only answered questions prior to the video fragment. It should be noted that responses from group 1 were categorized as "retrieval mechanism" responses whenever they explained that pop-up questions helped to check understanding or think about the content. As the answers were categorized blindly, it was not clear whether student referred to their thinking about either the previous video fragment (retrieval mechanism) or their prior knowledge (activation mechanism). It is likely that responses from group 1 categorized as "retrieval mechanism" actually refer to the activation of their prior knowledge, simply because pre-questions do not reflect upon previous video fragments. Some students clearly explained how questions activated their prior knowledge, and this mechanism was reported more often by students of group 1 (25%) than students of group 2 (1%) and group 3 (0%). Instead, students of group 2 (16%) reported more often how the questions signaled them to focus on certain topics of the video than students of group 1 (0%) and 3 (1%). Only few students reported that the questions encouraged them to review the video content.

Students were similarly asked to explain what they did not like about the pop-up questions. None of their responses were related to learning mechanisms but either focused on technical issues or on the format of the pop-up questions used **(Supplementary S1)**. Remarkably, students that answered pre-questions in group 1 (59%) mentioned more often that they did not like the timing of the pop-up questions than students of group 2 (3%) and group 3 (2%). This result is in agreement with interview responses of experiment 1, wherein students of group 1 explain that they believe pre-questions are not contributing to learning.

Test scores

In the final study we aimed to investigate whether students' ability to memorize the corresponding video content was also different for students that received either pre-questions and/or post-questions. Students' understanding of the video content was tested with a test containing questions on the same concepts as the pop-up questions from the videos. The test was performed individually and under supervision of a teacher. The test was scored with points from 0 to 12. The results reveal that the test scores were similar for students of both group 1 ($M=8.8$, $SD=1.8$), group 2 ($M=8.9$, $SD=1.57$) and group 3 ($M=9.0$, $SD=1.8$), $F(2, 224) = 0.16$, $p = 0.85$, $\omega 2=-0.01$ **(Figure 3)**.



▲**Figure 3** Mean test scores of pop-up questions within the video and of a test after the video for students in group 1 (pre-questions, n=86 and n=43), group 2 (pre- and post-questions, n=74 and n=86) and group 3 (post-questions, n=84 and n=98). The maximum possible score of both pop-up questions and test was 12 points. Error bars represent standard deviations of the mean. NS indicates no significance in test scores between the groups.

Discussion

The current study explored if and how pre-questions and post-questions within educational videos have different effects on learning. Both pre-questions and post-questions are known to improve learning from video (Carpenter and Toftness 2017; Lavigne and Risko 2018; Szpunar et al. 2013; Toftness et al. 2018; Vural 2013). However, the learning effects of these two types of questions are expected to evolve from different learning mechanisms since pre-questions preview upcoming information and post-questions review previous information. In this study we explored whether different types of pop-up questions within video have different effects on learning and how students believe that these different questions promote their learning.

The results of this study showed that students of all groups recognize a motivation mechanism by explaining that pop-up questions increase their focus of attention. The results of our previous study already suggest that post-questions increase students' focus since students that answered these questions were less likely to zapp back and forth through a video (Haagsman et al, 2020). Other studies on video lectures also show that students report less mind wandering when post-questions are present (Jing et al 2016; Szpunar et al 2013). We expected that the motivation mechanism is also an important learning mechanism of pre-questions since pre-questions give a sneak preview of the upcoming information and are hence expected to promote focus and interest in the next video fragment. The student responses in this study indeed suggest that both post-questions and pre-questions result in a higher focus of attention on the video content. It is assumed that all learning processes require a certain focus of attention, making the motivation mechanism as one of the most important mechanisms of learning (Schweppe and Rummer, 2014). The increased focus of attention perceived by students of all groups thus advocates the use of pop-up questions in general.

Another mechanism that is frequently reported by students of all groups is that pop-up questions help to test understanding and think about the video content. Both of these processes require retrieval of information from the memory and this retrieval mechanism is thought to promote retention (Carpenter et al. 2008; Chan et al. 2006; Karpicke 2009; Karpicke and Blunt 2011; Karpicke and Roediger 2007). It should however be noted that students that answer pre-questions cannot test understanding from a video fragment they have not yet seen. Thus, student responses referring to the retrieval mechanism in the pre-question group do not indicate retrieval of previous video content, but conceivably retrieval of their prior knowledge. Unfortunately, this difference could not be made from blinded student responses that merely explain that questions help to *"think about the content"*. In any case, the added value of pop-up questions as tools that promote thinking and test understanding is commonly acknowledged by students of all groups. Remarkably, only few students report that the pop-up questions promote reviewing of a specific topic. This does not necessarily mean that pop-up questions do not promote reviewing mechanisms. In fact, in a previous study we showed that students review and rewind the video more often just after post-questions appear (Haagsman et al. 2020). It should be noted that students in that study watched the videos at home, whereas students in the present study watched the videos in a controlled setting under supervision of a teacher. This controlled setting might discourage students to reread a textbook or rewind a videofragment, which would explain why students did not frequently report reviewing mechanisms in this study.

Interestingly, the present study also shows that students award different additional benefits to pre-questions, post-questions or a combination of both. The different perceptions suggest that other learning mechanisms are indeed at play. Students that answered only pre-questions (group 1) explained that they felt anxious because they could not always answer pre-questions with their prior knowledge, although they claim more often that pop-up questions help to activate their prior knowledge on that topic. This belief is in alignment with the proposed activation mechanism, which is thought to result in better integration of new and known information. Students that read questions before but answered them after the corresponding video content (group 2), were instead more positive towards pre-questions and explained that pre-questions clarify which video content is most relevant. This notion refers to the proposed signaling mechanism, which is thought to increase focus on the relevant video content. It is remarkable that students report less anxiety and more signaling events once they know they have to answer these questions after the video fragment. These results suggest promising learning mechanisms of combinations of pre- and post-questions, and we encourage further studies that examine whether students that answer pre-questions after a video indeed have higher focus on the related topics than students that answer these pre-questions right away.

The current study shows that students at least have different perceptions on the benefits of pop-up questions when answering them either before and/or after its corresponding video fragment. Nonetheless, no differences are shown in test performances for students that answered either pre-questions, post-questions or a combination of both. Thus, even though students' satisfaction and learning mechanisms might be different, one question type does not outperform the other on students' short-term retention and understanding. Of note, we believe that the tests in this study give a good representation of students' understanding of the pop-up questions since their level and content were alike. One explanation for the similar test scores between the three groups might be that a higher focus, the motivation mechanism suggested by students in all groups, is the main mechanism that improves retention and understanding from video. As already discussed, focus of attention is thought to be a prerequisite for learning and pop-up questions indeed appear to increase students' focus (Schacter and Szpunar 2015; Schweppe and Rummer 2014). A second explanation might be that, assuming that the reported learning mechanisms indeed occurred, previewing and reviewing mechanisms have similar effects on learning from video. We call for additional studies to examine whether the mechanisms that students report indeed occur and how these mechanisms affect retention and understanding from video.

This study is of explorative nature and provides information on how pre-questions and/or post-questions might have different mechanisms on learning from video. To our knowledge, this is the first study that compares the learning effects of pre-questions and post-questions in the context of video. Since previous studies reveal testing effects of pop-up questions, we advise teachers to include pop-up questions in their videos. Nonetheless, the current study shows that pre-questions and/or post-questions do not appear to have different effects on short term retention. In addition, student responses suggest that all of these types of questions promote a higher focus of attention, which is thought to be an essential mechanism of learning. Thus, we do not have a general recommendation on whether these pop-up questions should be posted as either pre-questions, post-questions or a combination of both. If the main aim of the questions is to stimulate students to process the video content and test understanding, then we recommend post-questions. If the purpose of the questions is to activate prior knowledge, then we recommend pre-questions. However, answering pre-questions can be frustrating, and from that perspective we would rather recommend teachers to combine these pre-questions with post-questions. We would also recommend a combination of pre- and post-questions if the aim of a question is to signal where students should focus on.

Informed consent and statement of human rights

Students were informed about the aims of the study and only participants that provided consent were included in the datasets of this study. The study is in agreement with the Review Board of Review Board of Social Sciences at Utrecht University (IRB approval number FETC 180-962).

Supplemental tables and figures

▼**Supplementary S1** Number of students that report shortcomings for the pop-up questions for students of group 1 (pre-questions), group 2 (pre- and post-questions) and group 3 (post-questions). Note that some students reported more than one shortcoming.

Points of improvement	Example	Group 1 (n=78)	Group 2 (n=57)	Group 3 (n=62)
Less interruption	"It is annoying that the video is suddenly stopped."	2 (3%)	9 (16%)	8 (13%)
Higher number	"There should be some more questions so that there is one question for all important concepts ."	7 (9%)	9 (16%)	17 (27%)
Other level	"The questions were rather easy in this case."	10 (13%)	15 (26%)	13 (21%)
Other content	"Could sometimes be more focused on the most important information."	2 (3%)	2 (4%)	1 (2%)
Less mistakes	"For one question the answer was marked as incorrect, while the feedback explained it was correct"	8 (10%)	0 (0%)	1 (2%)
Better formulation	"Sometimes the wording is a bit crooked."	1 (1%)	1 (2%)	1 (2%)
Different timing	"I would rather have the question posted after the belonging information, and not before."	46 (59%)	3 (5%)	1 (2%)
Unnecessary preview	"For me it is not necessary to announce the pop-up question, then you pay more attention from yourself."	0 (0%)	4 (7%)	0 (0%)
More feedback	"Sometimes it is not clear why something is good or wrong"	4 (5%)	4 (7%)	10 (16%)
Technical issues	Negative remarks related to general technical issues of the video platform	13 (17%)	18 (31%)	17 (27%)

3

4

Be Prepared! How pre-lab modules affect students' understanding of gene mapping

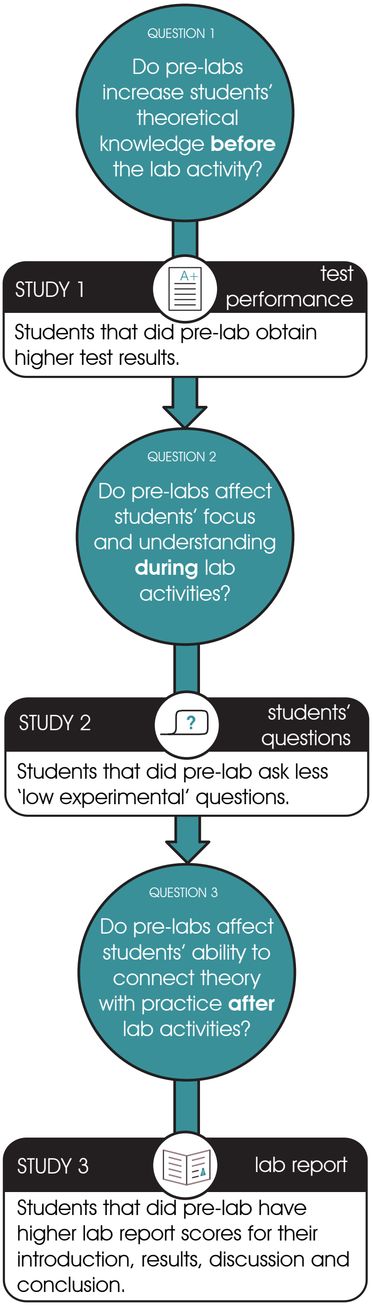
Marjolein E. Haagsman[1], Margot C. Koster[1], Johannes Boonstra[1], Karin Scager[2]

based on

Journal of Science Education and Technology 2021, 30, 461-470

1Department of Biology, Faculty of Science, Utrecht University

2 Department of Education, Faculty of Social and Behavioural Sciences, Utrecht University


QUESTION 1
Do pre-labs increase students' theoretical knowledge **before** the lab activity?
STUDY 1
test performance
Students that did pre-lab obtain higher test results.
QUESTION 2
Do pre-labs affect students' focus and understanding **during** lab activities?
STUDY 2
students' questions
Students that did pre-lab ask less 'low experimental' questions.
QUESTION 3
Do pre-labs affect students' ability to connect theory with practice **after** lab activities?
STUDY 3
lab report
Students that did pre-lab have higher lab report scores for their introduction, results, discussion and conclusion.

Summary

Lab activities are characteristic of life science education. In the current study, we investigate whether pre-lab modules can improve students' understanding of the theories and experimental procedures associated with lab activities. Such effects were studied in context of an expository lab on gene mapping in biology undergraduate education. An experimental group of 126 students had access to an online pre-lab module to prepare for the lab activity; a control group of 90 students did not have access to this pre-lab module. The data revealed that students who studied the pre-lab module had a better understanding of the gene mapping theory, at the onset of the class, when compared to the control group. Additionally, these students appeared to ask fewer questions on what needed to be done in the lab, suggesting more awareness of the experimental procedure. Further, students who studied the online pre-lab module showed greater understanding of the theory in their lab reports. These findings suggest that students' understanding of background theory and its relation to practice can readily be improved by enriching existing expository labs with pre-lab modules that contain information and questions on the complex conceptual information relevant to the lab experiment.

Introduction

Lab activities form a distinctive type of learning process within life science education. Indeed, they are widely considered to be essential for learning practical skills and investigative skills and grasping related theoretical concepts (Reid and Shah 2007). There exist various types of lab activities that differ in their suitability for meeting a given learning goal (Brownell and Kloser 2015). Although their importance is clear, studies on lab activities have shown that students regularly fail to gain intended learning aims—especially those related to understanding theoretical concepts (Hofstein and Lunetta 2004; Johnstone and Al-Shuaili 2001; Kirschner et al. 1993; Reid and Shah 2007).

The most common type of lab activity in large-scale undergraduate courses are expository labs or confirmatory labs in which research questions and methods are already set. Such classroom activities are often also referred to as "cookbook labs", given that, in such structures, students can follow preset procedures as a recipe, without understanding the discrete purpose of each step within the experimental procedure (Brownell and Kloser 2015; McComas 2005). Since both the experimental procedure and theoretical framework are already set, expository lab activities generally do not incorporate learning objectives such as posing research questions and designing methods to pursue them (Brownell and Kloser 2015; Domin 1999). Instead, they mainly stimulate students to develop practical skills, interpret data, and draw conclusions. In an explorative study we asked bachelor-level life science teachers to delineate the main objectives of their lab activities. The most frequently-mentioned learning objectives were to improve understanding of the theoretical framework and to relate this theory with practice—a set of goals that raise questions about the alignment of the learning goals and outcomes (unpublished results). However, studies that measure the efficacy of theoretical learning objectives in biology lab education still remain scarce. The present study aims to measure and improve these learning outcomes, applied to an expository lab activity on gene mapping.

The difficulty of improving theoretical understanding from lab activities is that students generally appear to be solely focused on the experimental procedure when working in the lab (Gunstone 1991; Hofstein and Lunetta 2004; Kozma 2003; van der Kolk et al. 2012). Support for this statement comes from the observation that students almost exclusively ask questions in the lab regarding the experimental procedure, (e.g. "*where can I find …?*" and *"when can I do …?"*: Kozma 2003; van der Kolk et al. 2012). This behavior may result from the limited working capacity of the brain and the large amount of new information they need to process in the lab (Johnstone et al. 1994). To clarify, students are required to recall information, make observations, search for materials, follow procedures, and analyze data, all at the same time. Thus, they are at risk of becoming cognitively overloaded—a condition that makes it more difficult to focus on the purpose and theoretical framework of the experiment (Johnstone 1997). This can result in inconsistencies between intended and actual learning outcomes with respect to theoretical understanding (Hofstein and Lunetta 2004; Johnstone and Al-Shuaili 2001; Reid and Shah 2007).

One solution to increasing the theoretical understanding, both of the experimental procedures and the relevant biological concepts, is to redesign expository labs into inquiry-based labs. In inquiry-based labs, students are required to design their own experiments; this can vary from setting up their own research question(s) to defining relevant theory and designing their own methodology (Zhang 2016). It is thought that this format will keep students more engaged and focused on the actual purpose of the experiment (Brownell and Kloser 2015). Previous meta-studies on inquiry-based labs (Furtak et al. 2012; Minner et al. 2010; Schroeder et al. 2007) have mostly revealed positive effects on students' learning outcomes. In addition, students in inquiry-based labs have appeared to ask more questions in the lab for which critical thinking is needed, suggesting that they better understand the purpose and theoretical framework of the lab activity (Hofstein et al. 2005).

Nonetheless, inquiry-based labs have some practical implications that suggest they may not be efficient or feasible for all levels and settings. The first implication is that students need to have some knowledge of lab techniques (Gormally et al. 2009; Krajcik et al. 2000). Besides, some experiments are simply too complex for students to design themselves (an issue that is particularly true for the gene mapping experiment explored in this study). An additional implication is that inquiry-based labs often appear to be more expensive and require more lab facilities and lab space (Gormally et al. 2009; Johnstone and Al-Shuaili 2001; Wei and Woodin 2011). Most importantly, inquiry-based labs focus on the entire process of conducting research—meaning that students generally need to formulate hypotheses, solve problems, apply theoretical knowledge, design experiments, use practical skills, select data, interpret data, derive conclusions, and identify limitations of the procedure (Johnstone and Al-Shuaili 2001). In other words, a great deal of time and effort is spent on objectives other than improving the understanding of biological concepts. Thus, inquiry-based labs may not be the best fit for lab activities designed to improve understanding of multiple biological concepts and procedures and taught to large groups of students with little lab experience.

An alternative method for increasing theoretical knowledge is to enrich expository labs with lab-preparation activities. It is hypothesized that more preparation will result in less cognitive overload in the lab and thus allow students to focus more on understanding the experiment (Johnstone 1997; Sweller et al. 1998). Better understanding can be readily obtained by handing out slides, videos, questions, and tests prior to the lab session(s) (Nadelson et al. 2015; Pogacnik and Cigic 2006; Whittle and Bickerdike 2015). A more modern method of preparing students for the lab is to use a computer module: a so-called pre-lab module.

Previous studies on pre-lab computer modules are mainly focused on the learning effects of experimental procedures (Johnstone 1997; Jones and Edwards 2010; Schmid and Yeung 2005). For example, previous pre-lab modules have been shown to increase students' confidence in doing dissection-based lab activities (Jones and Edwards 2010). Students also appear to perform better in the execution of chemical lab experiments when prepared with pre-labs (Johnstone 1997; Schmid and Yeung 2005). In addition, students who did pre-labs have been found to ask more theoretical questions (Winberg and Berg 2007) and less "thoughtless" questions on the experimental procedure that, according to the teachers, they could have easily answered themselves (Johnstone 1997).

Learning objectives

In the current study, we scrutinized a lab activity with high theoretical complexity that is designed for a large group of undergraduate students with scant lab experience. In this specific lab activity, students must approximate the genomic location of a certain gene that results in a visible phenotype when mutated. This exercise was chosen because the underlying theory on genetics, recombination, and gene mapping is difficult for undergraduates (Makarevitch & Kralich 2011). The learning objectives for this gene mapping experiment are that students can i) understand the biological principles important for gene mapping, ii) understand (the purpose of) the methods used during the experiment, iii) relate the theoretical knowledge of methods and biological principles to the research question, iv) perform the experiments, v) interpret the data and vi) identify the limitations of the experimental procedures.

We aimed to use pre-lab modules to improve the learning objectives on the understanding of the biological principles, the experimental procedure and of how this procedure is related to these principles. In other words, the present study addresses the following main question:

Do pre-lab modules improve students' understanding of the theory from lab activities with high theoretical complexity?

More specifically, we aimed to examine students' understanding of the experimental procedure, theory, and obtained data through all stages: before, during, and after the lab activity. The current chapter examines the following three specific questions:

1. *Do pre-lab modules increase theoretical knowledge before the lab activity, at the onset of the experiments?*
2. *Do pre-lab modules affect students' focus and understanding towards the methods, theory, and results during the lab activity, when doing experiments?*
3. *Do pre-lab modules affect students' understanding of the methods, theory, and results after the lab activity, when writing lab reports?*

The pre-lab considered in this study is designed to improve students' understanding of the theory behind a gene mapping experiment. The pre-lab includes videos, text, images, questions, and feedback on the theoretical background, experimental procedure, and interpretation of hypothetical data. Although the pre-lab studied herein is specifically designed for gene mapping, we expect the results of this study to be applicable to other lab activities with similar high theoretical complexity.

Methods

Participants

The participants in this study were students participating in the course, Molecular Genetic Research Techniques at Utrecht University. This course is one of many electives within the second and third years of the bachelor-level study in biology. Based on ethical considerations (namely to avoid unequal treatment of students within the same course) the research was conducted over two years. Correspondingly, data for the control group and experimental groups were collected across two consecutive years. One hundred twenty-one students participated during 2016—2017 (control group) and 149 students participated during 2017—2018 (experimental group). The course set-up and lab activities remained the same for both years. Students were asked to sign an informed consent containing information on the research and gathering of data of the study. Only participants providing informed consent were included within the datasets on descriptive statistics, pre-lab tests, and lab reports **(Table 1)**. Informed consent was provided by 90 students of the control group and 126 students of the experimental group. During the course, participants took an exam about topics unrelated to the one discussed in this study. The grade for this exam was used to compare the level of the control and experimental group.

▼**Table 1** Descriptive statistics of participants who prepared a lab activity on gene mapping either with or without a pre-lab module.

Group	Number of students	%male	%female	Average second exam grade[1]
Without pre-lab module	90	49%	51%	6.63 ± 1.52 (N=87)
With pre-lab module	126	40%	60%	6.71 ± 1.14 (N=104)

[1] The presented grade is the average grade for the exam of the second part of the course. This exam tested knowledge unrelated to the studied experiment in the first part of the course. Students are graded from 0.0 (lowest) to 10.0 (highest). Grades below 5.5 are considered insufficient. Note that some participants did not take the second exam.

Course design

The course is divided in four segments of two part-time weeks that are related to a specific research field; microbiology, molecular plant physiology, cell biology and developmental biology. Students have one optional lecture a week on the biological concepts important for the lab activities of that week. Each part contains mandatory expository lab activities guided by a specialized teacher in that specific field, of which none are (co)-authors of this study. The main teacher was usually assisted by two lab assistants: a master's student and Ph.D. student. The lab activities were taught in three groups of 30—40 students and experiments were performed in pairs. For each part of the course, lab activities were followed by a mandatory in-class discussion on the experiments and results. After each part, students were also required to complete a lab report on each of the experiments. Both the lab report and students' attitude in the laboratory were graded.

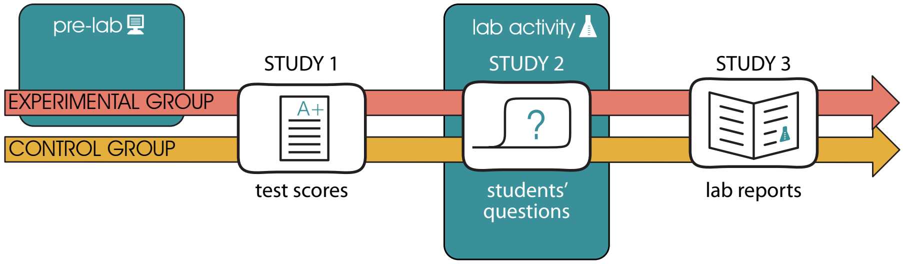
Lab activity on gene mapping

The study considered one of the expository lab activities during the part of molecular plant physiology and was performed in three separate sessions of two hours and 45 minutes. Students had to approach the genomic location of a mutation causing a certain phenotype in sandrocket (*Arabidopsis thaliana*). The location of the mutation was determined with a so-called gene mapping approach. During these exercises, students needed to isolate DNA from plant material, perform polymerase chain reactions (PCR) and separate PCR products with gel electrophoresis. This procedure was done for two parental plants and for their F2 offspring with mutant phenotypes. The sizes of each of the PCR products were compared between parents and F2 offspring to determine how often crossing over occurred between a specific location and the mutation (and thus if the mutation could be expected to be near that location). Information on the theory and experimental procedure of gene mapping was included in a lab manual. Both control and experimental groups were asked to read this lab manual before the onset of the lab activities.

Pre-lab module design

The experimental group was also required to complete two mandatory online computer modules (about 90 minutes) at home before the start of the lab experiment. These pre-lab modules were specifically designed for this gene mapping experiment and aligned with the first five (theoretical) learning objectives of this exercise on i) crossing-over, ii) method of gene mapping, iii) determining genotypes iv) calculating distances between genes and markers and v) reliability of gene mapping. The first module focused on gene mapping, which was also explained in the lab manual. The online module contained animations and short texts explaining homologous recombination, crossing over, and calculation of recombination frequencies. The information was alternated with questions on cross studies and gene mapping. This first module concluded with an explanation of the calculation of recombination frequencies needed for data analysis. The second module was mainly focused on the experimental protocol of gene mapping. This segment contained questions and explanations on the purpose of each protocol step. The module concluded with a possible outcome and questions on the interpretation of these data. The full protocol of gene mapping could be downloaded at the end of the second module. The modules were made with Xerte online toolkits (online software), exported as a SCORM 2004 3rd Ed package and uploaded in Blackboard (learning management system). Before the lab activity, students' use of the module was checked with records presented in Blackboard.

The effects of the pre-lab module on understanding the theory and experimental procedure were measured with a pre-lab test, annotations of student questions, and lab report scores. A schematic presentation of this set-up is shown in **Figure 1**.



▲**Figure 1** Schematic presentation of the three studies on pre-lab tests scores, student questions and lab report scores of the control and experimental group. The arrows present the sequence of activities and studies per group.

Pre-lab test

Students were asked to complete a theoretical test on gene mapping at the start of the gene mapping experiment. Students were not informed that they would receive this test on that particular day. It was explained that the test was only meant for research purposes and that students would not be graded for this test. Teachers did not receive individual test scores. The first two questions of the pre-lab test evaluated whether the student had read the lab manual and attended the lecture before the start of the lab activity. These items were followed by nine multiple-choice questions and true-false statements on the main principles of gene mapping. Each sub-question was scored with either one or zero points. The average test scores for the experimental and control groups were compared in IBM SPSS statistics Version 26.0 (statistics software) using an independent t-test. Possible additional effects from reading the lab manual or attending the lecture on test scores were analyzed using a stepwise multiple regression, with the experimental group as first predictor and reading the lab manual and attending the lecture as second predictor.

Student questions

Lab activities were taught in three classes of 30-50 students and students' questions were collected during the lab-activity. The main teacher was asked to clip a voice recorder on his lab coat and make recordings during all lab activities on gene mapping. The teacher was informed that these recordings were only intended to record students' questions and that his own explanations and instructions would not be analyzed. Recordings of only one of the three classes of each year were selected for analysis since other recordings were not complete, with 46 students in the control group and 39 students in the experimental group. All students' questions concerning the gene mapping experiment were transcribed except for questions that were repeated by the same student. Questions that were hard to decipher but clarified by the teachers' answer were still formulated and transcribed. The data were analyzed in several stages using Nvivo (qualitative data analysis computer software). Student questions were first coded in meaningful categories with an inductive coding approach (Creswell 2007). The categories were then discussed with the research team and merged, deleted or reformulated into new categories. These two steps were performed twice, after which agreement was reached (Thomas, 2006). The 386 transcribed questions were categorized in the following eight categories: general organization, theoretical low-order, theoretical high-order, experimental low-order, experimental high-order, interpretation low-order, interpretation high-order or other **(Table 2)**. Questions were categorized as low-order questions if they did not require understanding of the theory and/or experimental procedure. Questions were categorized as high order questions if they could only be asked with some understanding of the theory and experimental procedure. Questions on the background theory were defined as theoretical questions, questions on the experimental procedure as experimental questions and questions on the interpretation of the obtained results as interpretation questions. All questions on organization of the experiment, such as where to find or store materials, were defined as general organization questions. The assignations of the transcribed questions into these categories was performed blind by the first author (who is experienced in teaching biology). Afterwards, a selection of annotations was checked blind by the second, third, and fourth authors. Annotations that were disagreed upon were discussed to reach consensus.

Lab reports

Each student pair was required to complete a joint digital lab report after the final in-class discussion. No interim feedback was provided on the lab reports and the reports were required to include an introduction, methods, results,

▼**Table 2** Categorization of students' questions within the lab

Question category	Explanation	Examples
General organization	Questions on the general organization of the experiment	- Where can I find ethanol? - Where should I bring my sample? - Can we get some new solution?
Theoretical low order	Low-order questions on the theoretical background of the experiment	- What do you mean with ecotype? - What do you mean by Landsberger; a wildtype plant?
Theoretical high order	High-order questions on the theoretical background of the experiment	- So the mutation is nearby if they are homozygous for all mutants? - Do you first search for primers complementary to the mutants because they may not be complementary to the wildtype?
Experimental low order	Low-order questions on the procedure of the experiment	- Do we also need to add loading buffer in here? - What voltage do we need?
Experimental high order	High-order questions on the procedure of the experiment	- Wouldn't it be better to add Taq polymerase after making those PCR thingies? - Something went wrong last time… so should we add more loading buffer?
Interpretation low-order	Low-order questions on the interpretation of the results of the experiment	- Why did it fail? - What can I interpret; Just whether it's homozygous or heterozygous?
Interpretation high-order	High-order questions on the interpretation of the results of the experiment	- Is it possible that the primers were not mixed well enough? - But the parents couldn't be heterozygous, right? Because it seems that they are…?
Other	Every question that does not fit in any other category	- How expensive is one such tube?

conclusion and discussion. Each of these five lab report sections were aligned with the first (comprehensive) learning objectives of the gene mapping experiment as students were expected to i) understand the biological principles important for gene mapping, ii) understand the methods used during the experiment , iii) relate the theoretical knowledge of methods and biological principles to the research question, iv) perform the experiments, v) interpret the data and vi) identify the limitations of the experimental procedures.

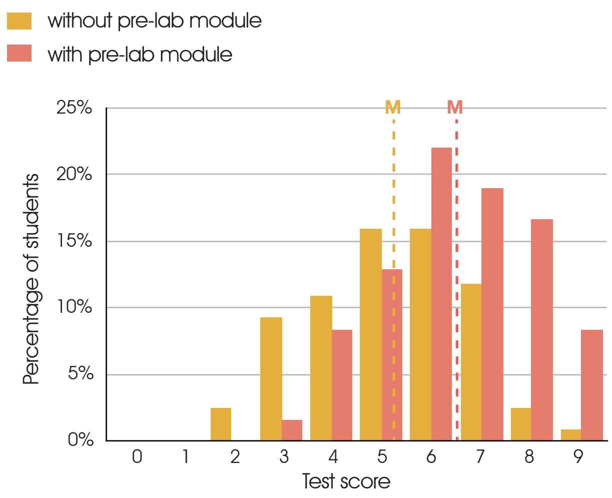
All 91 lab reports were anonymized; and dates revealing whether students belonged to the experimental or control group were removed from the files. The reports of both the control and experimental group were given random numbers and were blindly assessed by two examiners (second and third co-authors). A simple rubric was used to assess the reports on introduction, methods, results, conclusions, and discussion with either an insufficient, sufficient, or excellent categorization. Note that the rubric used in this study was a modified version of the actual rubric used in the course and was only created for the purpose of this study, merely focusing on students' understanding of the experimental steps and theory behind gene mapping. The actual rubric used for feedback and grading of students is more elaborate and also focuses on general content, lay-out, relevance and completion. Report sections were not assessed if students left them entirely blank.

The examiners first assessed 10 reports and then came together to discuss their grading system. The rubric was fine-tuned and used by both examiners to assess the reports in three phases. First, examiners assessed all reports individually and equally scored 71% of all rubric categories equally which yielded a Cohen's kappa of 0.55, suggesting moderate agreement (Landis and Koch 1977). Second, the examiners discussed their grading for reports they assessed differently and individually assessed these reports again. Third, final discrepancies in grading were discussed to arrive at a final consensus on the grading of each rubric category for each report. The 2x3 contingency tables of rubric scores (insufficient, sufficient or excellent) of the control and experimental group were compared with a Fisher Exact Test in IBM Statistics Version 26 (statistics software).

Results

Pre-lab test scores

The aim of the pre-lab modules was to increase students' knowledge in advance of the lab exercise in an attempt to reduce the cognitive load on students and connect the theory to the lab experiment. To test the assumption that the module indeed increases theoretical knowledge, a test was performed at the start of the experiment. The test was scored from 0 (lowest) to 9 (highest) **(Figure 2)**. The students who studied the pre-lab module indeed scored significantly higher (M=6.5, SE=0.14) than the control group (M=5.2, SE=0.17) ($t(197) = -6.10$, $p<.001$). Thus, this result demonstrates that students have more theoretical knowledge regarding the experiment at the start of the lab activity when prepared with the pre-lab modules.



▲**Figure 2** Test scores on gene mapping for students who did (N=116) or did not (N=83) prepare with a pre-lab module. The theoretical test was performed at the start of the mapping experiment. Students received scores from 0 (lowest) to 9 (highest). The dashed lines represent the mean adjusted test scores for the control and experimental group.

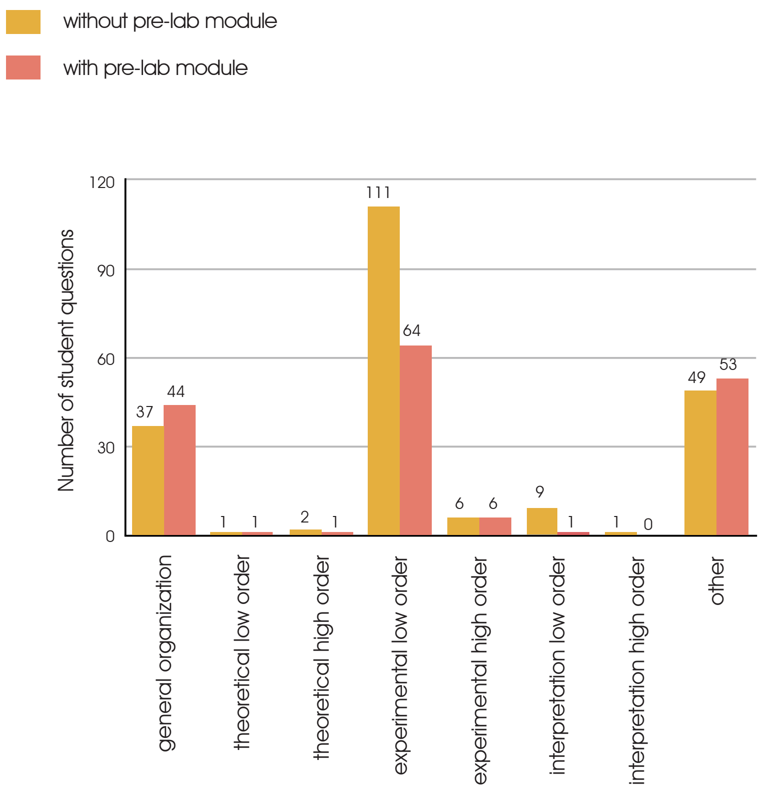
The theoretical background information on gene mapping was also discussed in a lecture and included in the lab manual. The aim of the lab manual was similar to the pre-lab modules: to improve understanding of the theoretical background and method of gene mapping. Both the control and experimental groups were asked to read the lab manual before start of the lab activities, but the main teacher had strong doubts as to whether students actually read the lab manual prior to the lab. We therefore included a question on reading the lab manual in the pre-lab test. Only 47% of the experimental group and 66% of the control group answered this question. Interestingly, these students claimed to have read the lab manual significantly more often in the experimental group (56%) than in the control group (13%) (χ^2 (1, N=110) = 23.158, p<.001) **(Supplementary S1)**. Nonetheless, reading the lab manual does not have significant additional effect on the variance in test scores **(Supplementary S3)**. Similarly, no significant additional effect on the variance in test scores is found in relation to lecture attendance **(Supplementary S2 and S3)**.

Students' questions during lab activities

Teachers of the present course stated that students usually almost exclusively ask what they need to do at that specific moment of the lab activity. If students have more prior knowledge on the theoretical background of a lab activity, it is expected that students are better aware of what needs to be done during the lab exercise and thus ask fewer of these types of questions. Similarly, it is expected that these students will change their focus towards the actual purpose of the experiment. These hypotheses were tested by recording, transcribing, and annotating student questions within the lab for one class of both the control group and experimental groups **(Table 2)**.

The recordings show that only 4% of all student questions are high-order questions that display comprehension of the theory, experimental procedure, or data **(Figure 3)**. In fact, 111 out of 179 student questions in the control group were questions on what needs to be done during the lab activity. Remarkably, students in the experimental group only asked 64 of such experimental low-order questions. Similarly, low-order questions on how to interpret the data are asked only once by the experimental group, compared to nine times by the control group. Questions on the general organization such as where to find or bring lab material are asked roughly as often for both groups.

In summary, the number of high-order questions is relatively low for both the experimental and control groups. Nevertheless, the number of low-order questions on the experimental procedure is lower for students who used the pre-lab module for preparation. More specifically, students in the control group asked 1.7 times as often what they needed to do at certain moments during the lab activity.



▲**Figure 3** Number of student question types asked to the main teacher during the lab activity for students prepared with (N=46) and without (N=39) a pre-lab module.

Lab report scores

The previous results show effects of pre-lab modules at the onset and during the lab activity. We also determined whether students were better able to process the theory, procedure, and results during a later stage when reporting their experiment. The students in the course were asked to write a lab report that was expected to contain an introduction, methods, results, conclusions and discussion. Each of these sections were expected to reflect one of the main

▼**Table 3** Percentage of lab report scores for students who prepared the gene mapping experiment either with or without a pre-lab module

	Without pre-lab module				With pre-lab module						
Criterion	N	Insufficient	Sufficient	Excellent	N	Insufficient	Sufficient	Excellent	FE	df	p
Introduction	37	11%	81%	8%	52	10%	42%	48%	17.76	2	<0.001***
Method	34	3%	41%	56%	52	0%	38%	62%	1.61	2	0.546
Results	37	32%	41%	27%	52	2%	35%	63%	20.35	2	<0.001***
Conclusions	26	38%	38%	23%	52	6%	31%	63%	16.52	2	<0.001***
Discussion	26	65%	23%	12%	52	35%	29%	37%	7.643	2	0.023*

Note. The number of scores differs per lab report criterion; some lab report sections were left blank by students.
*$p<0.05$ **$p<0.01$ ***$p<0.001$

theoretical learning objectives of the gene mapping experiment. The sections were scored blindly for both the control and experimental groups using a simple rubric to specifically assess students' level of understanding. Students who received the pre-lab module had significantly higher scores for four of the five lab report sections; only the Method section was scored similar for both the control and experimental group **(Table 3)**. Thus, students who had the pre-lab module were found to be better at connecting the lab activity to the theory (i.e., presenting the theory needed to clarify the research question, presenting the results needed to answer the research question, interpreting the results correctly and proposing future experiments).

Discussion

The main purpose of this study was to increase students' understanding of the complex theory of gene mapping and its relation to practice. The investigation demonstrates that pre-lab modules can improve students' understanding of gene mapping during an expository lab activity. More specifically, students who studied the pre-lab module showed better understanding on the gene mapping theory at the start of the experiment, as compared to those who had not received that module. Secondly, students who studied the pre-lab module posed few questions about what they were required to do in the lab. Thirdly, lab report scores revealed that students who studied the pre-lab module could better relate the background theory with the research aim, select the relevant results needed to answer the research question, understand how the obtained results were related to the background theory, and understand the limitations of the data analysis and procedure. In other words, students who studied the pre-lab module knew better what to do in the lab and could better connect the background theory with practice.

Limitations

It should be noted that the data of the control group were collected one year prior to collecting the data of the experimental group. This has possibly led to subtle differences in lab instruction between the control and experimental groups. The main teacher has indicated that they set stricter deadlines for the lab reports in the second year of this study, which explains the small sample sizes of the control group for the analysis of lab reports. Nonetheless, other lab manual instructions, the main teacher, the lecture on gene mapping, and the setup of the experiments remained identical for the two groups. No data are available to investigate that the experimental and control group were comparable prior to this study.

Better preparation

It is interesting to find out that only a minority of students in the control group appear to have read the lab manual before the lab-activity. This is in agreement with previous research of Jones and Edwards (2010) showing that, when pre-lab modules are not used, only 15% of biology students claim to do a substantial amount of preparation before they enter the lab. In our study, viewing the pre-lab modules was mandatory; it is not certain the same results would be obtained if use of the modules had been optional. However, it has been shown that students prefer to prepare with computer modules with animations and tests than to read long texts such as from lab manuals (Bouwmeester et al. 2016; Jones and Edwards 2010). Moreover, online modules facilitate the monitoring of students' activities enabling teachers to make the pre-lab modules mandatory if needed. Another expected advantage of monitoring students' activity is that teachers could adapt their lab instructions to the individual needs of the students. Besides, interactive questions in pre-lab modules required students to stay more actively involved with the theory. It is likely that the alternation of such questions with theory increased students' focus and understanding of the presented (Haagsman et al. 2020).

It should be recalled that students who studied the pre-lab module claimed to have read the lab manual more than four times as often than the control group. Since other factors have remained the same, we hypothesize that the increased reading might actually be an effect of doing the pre-lab module. A possible explanation is that students are indeed more actively involved and that the pre-lab module might have triggered additional reading of the manual. Another explanation is that students might look for information in the manual to answer the assignments in the pre-lab module. One way or another, this result raises the question whether the presented results are direct effects of the pre-lab module itself or indirect effects of just reading the lab manual. Nevertheless, we found no significant additional effect of reading the lab manual on the test scores between the groups. This result implies that the improved results on students' test scores were not an effect of the lab manual. We have no reason to believe that this is different for the presented lab report scores or students' questions.

Improving understanding

In this study, we systematically explored students' insight into the experimental procedure by recording all questions from students and categorizing them in different question types. In general, students most regularly asked simply what they needed to do in the lab. This is consistent with previous studies on student questions in the lab (Johnstone 1997; Kozma 2003; van der Kolk et al. 2012). Most importantly, we showed that these types of low-order questions are asked less often by students who studied the pre-lab module. One advantage of fewer students' questions is that it reduces the workload for teachers and teaching assistants in the lab. This is especially relevant for labs with few assistants per students. Moreover, fewer low-order experimental questions suggests that students know better what needs to be done in the lab and have more confidence in that knowledge—a finding in agreement with previous results on pre-lab materials (Johnstone 1997; Jones and Edwards, 2010; Schmid and Yeung 2005).

However, the number of questions students ask on the background theory in our study remains scarce for both students with and without pre-labs. This result may be caused by the fact that there was no time scheduled for discussion during the lab activity and students are too occupied with the experimental procedures to reflect on the theoretical background. However, despite the fact that the number of theoretical questions did not increase in the experimental group compared to the control group, the theoretical understanding appeared to be clearly improved in the experimental group. The students that studied the pre-lab module scored better on the lab journal assessment, including their ability to discuss the results and to indicate limitations.

Recommendations and Future Studies

This study shows that the introduction of pre-lab modules results in better performance in a lab report and fewer low-order experimental questions in the lab, which we expect to be the result of improved understanding of the link between theory and practice. However, since reflecting on the theory during the lab exercises appears to be limited, we recommend that teachers use extra time in the lab to ask scaffolding questions and even further improve students' understanding of the experiment. Post-lab activities such as in-class discussions or post-lab computer modules might, similar to pre-lab modules, guide students to reflect upon their obtained results and improve students' understanding of the experiment (Reid and Shah 2007). We support the conduction of studies on the effect of such post-lab modules on students' understanding of the background theory and its relation to practice. Finally, we should note that this study was specifically performed to align lab activities with learning goals on understanding the theory and its relation to practice. We highly recommend that teachers and researchers investigate how to align their lab activities with their learning objectives.

This study started from an exploratory investigation showing that the main aim of lab activities of most undergraduate life science teachers is to improve students' understanding of background theory and its relation to practice. Students' understanding of the theoretical framework can be improved with inquiry-based labs, but such activities are not always feasible when multiple concepts are taught, lab space and time is limited, and students have limited lab experience. This study shows that students' theoretical understanding and its relation to practice can already simply be improved by enriching existing expository lab-activities with pre-lab computer modules.

Informed consent and statement of human rights

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee at Utrecht University (Review Board of Beta and Geosciences at Utrecht University, approval number S-19290). An informed and signed consent was obtained from all individual participants.

Supplemental tables and figures

▼**Supplementary S1** Descriptive statistics from pre-lab test scores for students that did or did not read the lab manual.

Group	Read lab manual	N	Mean	S.D
Without pre-lab module	Yes	7 (8%)	5.14	1.57
	No	48 (58%)	5.06	1.64
	Not determined	28 (34%)	5.43	1.43
With pre-lab module	Yes	31 (27%)	6.13	1.12
	No	24 (21%)	5.75	1.26
	Not determined	61 (53%)	7.02	1.54

Mean test scores are only shown for students that did or did not read the lab manual. The test was performed at the start of the mapping experiment. Students received scores from 0 (lowest) to 9 (highest). Pre-lab test scores of students that did not answer whether they read the lab manual are presented as 'not determined'.

▼**Supplementary S2** Descriptive statistics of pre-lab test scores for students that did or did not prepare with a pre-lab module and did or did not stated to attend or not attend the lecture before the lab activity.

4

Group	Attend lecture	N	Mean	S.D
Without pre-lab module	Yes	15 (18%)	4.60	1.88
	No	46 (55%)	5.24	1.42
	Not determined	22 (27%)	5.50	1.57
With pre-lab module	Yes	17 (15%)	6.18	1.29
	No	60 (52%)	6.42	1.49
	Not determined	39 (34%)	6.82	1.52

Mean test scores are only shown for students that did or did not attend the lecture. The test was performed at the start of the mapping experiment. Students received scores from 0 (lowest) to 9 (highest). One group of students could not attend the lecture because their lecture was scheduled after the pre-lab test and their test scores are presented together with other students that stated not to have attended the lecture. Pre-lab test scores from students that did not answer whether they attended the lecture are presented as 'not determined'.

▼**Supplementary S3** Hierarchical Regression results of pre-lab test scores for students that did or did not prepare with a pre-lab module and claimed to attend or not attend the lecture before the lab activity.

Variable	B	95% CI for *B*		SE *B*	β	R^2	Δ R^2
		LL	UL				
Step 1						0.09	0.09**
Constant	5.08	4.69	5.45	0.19			
Control vs Exp Group	0.87	0.33	1.41	0.27	0.30**		
Step 2						0.10	0.01
Constant	5.16	3.36	6.96	0.91			
Control vs Exp Group	0.78	0.17	1.39	0.31	0.26*		
Read manual	-0.23	-0.88	0.43	0.33	-0.07		
Attend lecture	0.20	-0.42	0.81	0.31	0.06		

Note. CI = confidence interval; LL= lower limit, UL = upper limit; Contr Group = control group; Exp Group = experimental group
*p<0.05 **p<0.01 ***p<0.001

4

5

Examiners' use of rubric criteria for grading bachelor theses

Marjolein E. Haagsman[1], Basten Snoek[1], Anton Peeters,
Karin Scager[2], Frans Prins[2]

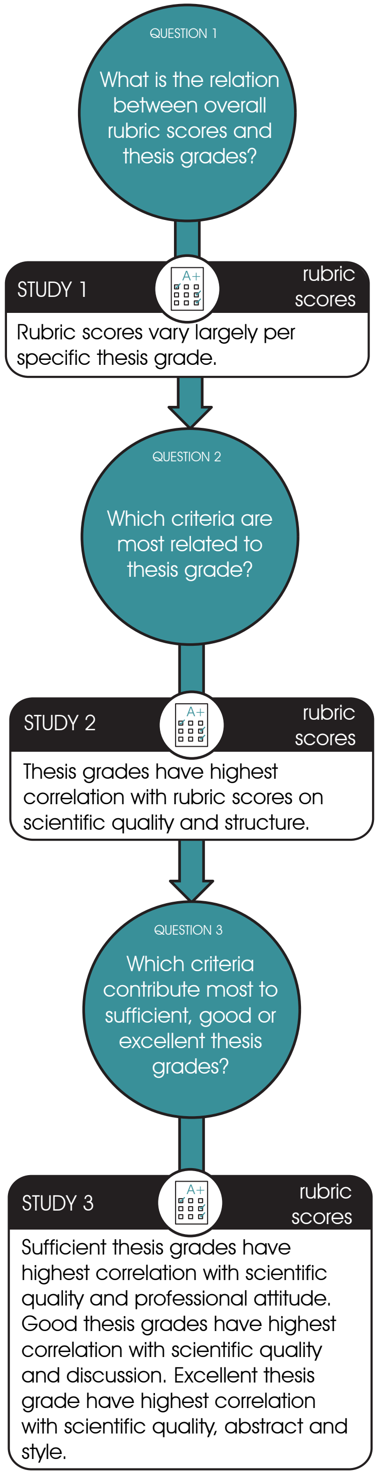
based on

Assessment and Evaluation in Higher Education, 2021

DOI: 10.1080/02602938.2020.1864287

1Department of Biology, Faculty of Science, Utrecht University

2 Department of Education, Faculty of Social and Behavioural Sciences, Utrecht University


QUESTION 1
What is the relation between overall rubric scores and thesis grades?
STUDY 1
A+
rubric scores
Rubric scores vary largely per specific thesis grade.
QUESTION 2
Which criteria are most related to thesis grade?
STUDY 2
A+
rubric scores
Thesis grades have highest correlation with rubric scores on scientific quality and structure.
QUESTION 3
Which criteria contribute most to sufficient, good or excellent thesis grades?
STUDY 3
A+
rubric scores
Sufficient thesis grades have highest correlation with scientific quality and professional attitude. Good thesis grades have highest correlation with scientific quality and discussion. Excellent thesis grade have highest correlation with scientific quality, abstract and style.

Summary

Students are generally required to demonstrate diverse skills when writing their bachelor thesis. Accordingly, examiners are expected to consider all these skills when assessing the thesis, regularly with one overall grade. In this study, we examine which criteria of a rubric contribute most to the overall assessment. The study is performed through quantitative analyses of 318 theses of undergraduate biology students. The analyses demonstrate that all criteria scores are predictive, but that scientific quality and professional attitude give the best prediction of thesis grade, together with structure. The predictiveness of scientific quality and professional attitude correspond with the instructions given to examiners that these are important criteria to consider. Presentation-related criteria scores on writing skills and expressing catchy and justifying titles give the lowest prediction of grade. Moreover, this study uniquely identifies that some criteria appear more predictive for low grades than for high grades, with professional attitude being a good predictor for low grades and abstract being a good predictor for high grades. We recommend similar analyses for students to help them prioritize on the most relevant criteria, for supervisors to instruct students on these criteria, and for education managers to evaluate whether bachelor theses are assessed on the criteria they find most relevant.

Introduction

Undergraduate students in the Netherlands commonly complete their bachelor studies with a 'capstone' thesis. This bachelor graduation thesis is considered an important part of the curriculum as it requires students to find relevant literature, process complex information, use scientific reasoning, think critically and work individually. All these skills need to be considered by the examiner when assessing the overall thesis quality and students' professional attitude, often with only one single grade. Thus, the examiner is required to interpret the quality of each assessment criterion and then needs to integrate all criteria to produce an overall quality score or grade. The current study explores how examiners weigh different rubric criteria in the overall assessment of undergraduate science theses. We quantified and analysed assessment criteria of the Biology undergraduate programme of Utrecht University, the Netherlands. Specifically, we aim to examine which criteria are considered most significant for the overall grade of research literature theses.

Quality criteria of bachelor theses

The bachelor thesis used for the current study consist of a review of existent primary research literature in the field of biology in the broadest sense. The main criteria for writing a good bachelor thesis in this study are to i) formulate a testable question, ii) select and process information from literature, iii) critically evaluate literature and iv) write a concise scientific report. In other words, students are required to complete the research cycle and show the belonging scientific skills. The steps of the research cycle are generally reflected in the required components of the thesis such as introduction, results and discussion (Prins, Kleijn and Tartwijk 2017). In the field of biology, Timmerman et al. (2011) developed a set of generic criteria to assess these components needed in the process of scientific writing. Criteria included were context and accuracy of the introduction, scientific merit and testability of the hypothesis, experimental design, data selection and presentation, analysis of results, significance and final conclusions of the discussion, alternative explanations and limitations of the discussion, use of primary literature and overall writing quality (Timmerman et al. 2011). Although these criteria were developed to assess reports in undergraduate biology laboratory courses, most criteria are applicable to every scientific writing assignment and also used in this study. Reddy and Andrade (2010) observed a gap in higher education research and endorsed more studies on the correlation of the quality of these criteria and final grading. To our knowledge, only few studies explored which criteria are of prime relevance when grading scientific writing assignments or bachelor theses. One study in PhD and master education shows that thesis grades are mainly related with the assessment of the research set-up and interpretation of obtained results (Bourke and Holbrook 2013). These criteria are however irrelevant for bachelors' theses that are solely based on literature research and do not require students to design a research set-up and obtain data. Another study on bachelors' theses in nursing found that thesis grades mainly correlate with student's ability to compare the reported studies (Lundgren, Halvarsson and Robertsson 2008). The theses described by Lundgren, Halvarsson and Robertsson (2008) are however not assessed with rubrics describing quality levels for every criterion. Here, we specifically assess the relationship of bachelor thesis grades and criteria scores that are acquired with the aid of such rubrics.

Goals of using rubrics from the student and examiners' perspective

Rubrics gained popularity in higher education over the last thirty years and are commonly referred to as 'a document that articulates the expectations for an assignment by listing the criteria, or what counts, and describing levels of quality from excellent to poor' (Reddy and Andrade 2010, p. 1). In other words, rubrics define criteria of a certain task and/or expectation and contain typifying descriptions of various quality levels for each of these criteria. The examiner uses the rubric to address the quality levels that fits best with the students' work. If well designed, rubrics help the students to assess their own work and direct them in what to improve for upcoming drafts or tasks (Tai et al. 2018). Furthermore, rubrics are designed to make students aware of the expectations of the examiners, which likely explains why rubrics have shown to reduce students' anxiety towards assignments (Andrade and Du 2005; Panadero, Alonso-Topia and Recke 2013). On the other hand, rubrics guide teachers and examiners in providing relevant feedback (Prins, Kleijn and Tartwijk 2017). From a governance viewpoint, rubrics can be administered alongside the graded final thesis, which facilitates (external) committees like the board of examiners and accreditation panels in validating the quality of the product, and curriculum as a whole. In this study we focus on how examiners use rubrics for assigning grades for summative assessment.

Rubrics for summative assessment

The summative assessment of theses is a complex process since many criteria are involved (Sadler 2009). Besides, there is not one single answer or best approach for writing such intricate assignments (Sadler 2009). In this study, examiners are provided with a rubric to guide them in the complex process of assessing bachelor graduation theses. The examiners can use the rubric to review and assess the distinct criteria, but no formula is provided to calculate final grades from these criteria scores. Thus, the examiners are free in how to use the rubric for determining the final overall grade.

It is logically relevant that these weighed grades accurately reflect the quality of students' work. Students in higher education greatly value the grades they obtain, as they do not perceive them as merely feedback instruments, but also as measurements of academic achievement needed for allowance to MSc programmes and even their future career (Goulden and Griffin 1995). This is especially true for the bachelor theses. Grades are however influenced by the experience and skills of the examiner, even when using rubrics with well-defined criteria (Kapborg and Berterö 2002). Besides, different examiners might put different 'weights' to each rubric criterion (Sadler 2009). Hence, it is mainly the examiner that determines the validity of a grade and not the instrument itself (van der Vleuten et al. 2012). This is especially an issue for the bachelor theses within this study, since these are assessed by a large variety of examiners from PhD students to full professors and from both within and outside the university and across biological disciplines, from ethics to molecular cell biology and from marine ecology to bioinformatics. Thus, we aim to improve insight into how thesis grades are assigned by examiners.

The rubric used in this study contains categories based on commonly required components and criteria of science reports, including; title, summary, introduction, set-up, discussion and conclusion, scientific quality, spelling and grammar, style, length and layout, figures and tables, references and professional attitude. Each criterion includes descriptions of subcriteria relevant for literature studies and comparable to the subcriteria described by Timmerman et al. (2011). Here, we explore how examiners weigh these criteria to decide upon the overall thesis grade. More specifically, we aim to address the following questions:

1. *What is the relation between criteria scores and thesis grade?*
2. *Which criteria are most predictive for grades of bachelor theses?*

This study is performed through quantitative analyses of criteria scores and thesis grades of 318 bachelor theses of three subsequent years in our Utrecht University Biology curriculum.

Materials and Methods

Participants

The rubrics included in our dataset were used to assess 318 students that wrote their thesis within the broad Biology bachelor programme of Utrecht University, the Netherlands. Rubric scores are only documented after students pass the course (thesis grade of 5.5 or above, scale 1.0 to 10.0). The theses evaluated were from all students that started their thesis project in the academic years; 2014-2015, 2015-2016 and 2016-2017 (14-15: 129, 15-16; 138; 16-17: 51). Of these, 156 (49.1%) were written by males and 161 (50.9%) by females. Henceforth, only the rubric assessments and thesis grades were collected and anonymously included in this study. The rubric assessments and thesis grade are always agreed on by two independent examiners, of which at least one is an Utrecht University Biology faculty member or approved faculty member of an affiliated department within the university and certified with at least a Basic Teaching Qualification. In total, 202 individual examiners were involved in grading. Of all theses, 12.3% were graded by external supervisors from e.g. companies, institutes or NGO's and had no direct affiliation with Utrecht University. Of all theses, 0.2% was graded by a technician, 5.2% were graded by a PhD student, 2.7% by 'teachers', 45.3% by assistant professors, 12.5% by associate professors and 17.1% by full professors.

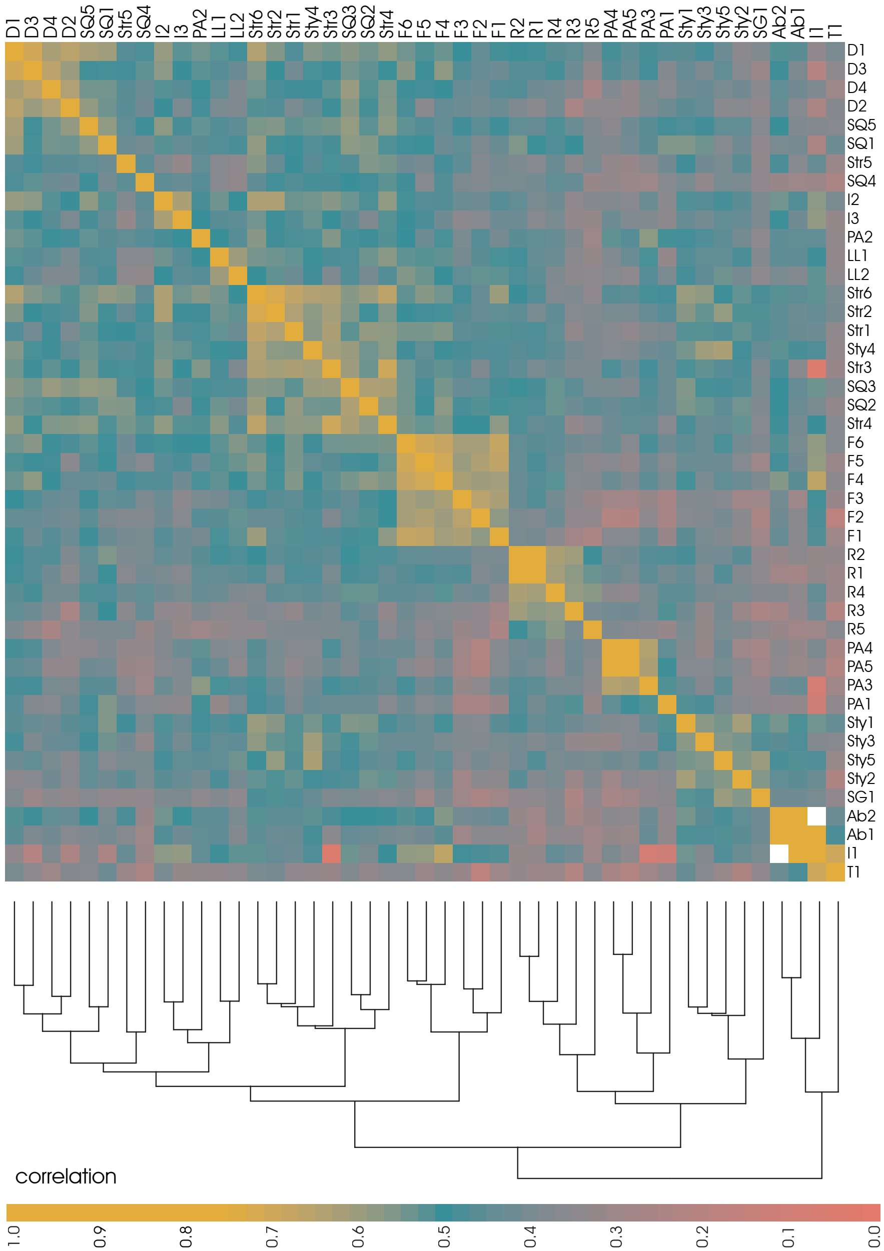
Context

It is compulsory for every Utrecht University Biology bachelor student to write a theoretical bachelor thesis that addresses a scientific question that falls within one of the many biological disciplines, including associated disciplines such as didactics, ethics, (sustainability) policy and consultation. The thesis should contain a well-defined research question that can be answered using mostly primary peer-reviewed scientific literature. The product is a well thought-over text of about 6,000-8,000 words (quality is leading over length) and may be written in English or in Dutch, with at least 20 primary sources cited, but usually many more are used. Next to the compulsory thesis, the majority of students also conducted a research project in the field or laboratory on the same or similar topic.

Course setting

Students spend 10 weeks half-time or 5 weeks full time on the writing process (total 200 hours). There are three compulsory seminars. In the first week an introduction lecture is organized where rules, guidelines and expectations are explained. During this session, the students are adviced to look at an online version of the rubric to get a view on the expectations of the examiners (**Supplementary S1**). In a second meeting organized by the Utrecht University Library, the student practice with literature searches, correct literature usage and referencing. Lastly, in week 3 of the course, aspects of academic writing are trained. For the rest of the time, students work individually under direct guidance of their topical (daily) supervisor. It is compulsory to hand-in a thesis plan (raw version of main question and ideas), on a set deadline date (approximately week 3).

During the writing process, the student usually hands in two versions of their thesis: a concept version and a final version. The student receives feedback on both versions from the daily supervisor on the content, structure, effort and the progress. Examiners assess the final version of the thesis with one grade that considers all assessment criteria.



▲**Figure 1** Calculated Euclidean distance hierarchically clustering matrix, indicating correlations between the scores of all subcriteria of all 13 categories of all numerated rubrics. The color range in the legend indicates the Pearson correlations between the subcategories (red=Pearson correlation of 0.0, yellow=Pearson correlation of 1.0). The subcriteria are ordered based on the similarity in scores, also indicated by the dendogram on the left. The subcriteria are numbered according to the rubric in Supplemtary 1 (T = title, I=introduction, Ab=Abstract, Str = Structure, D =Discussion, SQ = Scientific Quality, SG = Spelling and Grammar, Sty= Style, F = Figures, R= References, PA = Professional Attitude). Hence, subcriteria that are scored similarly, such as the discussion overview and discussion line of thought, are directly connected in this dendogram. The dendogram is calculated based on Euclidean distance hierarchically clustering.

5

Rubric scores and grading

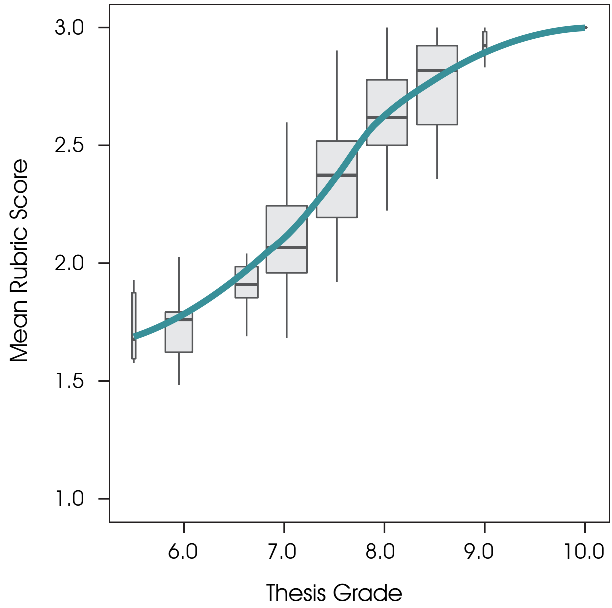
The final version of the thesis is assessed with the compulsory aid of a rubric. The rubric was developed based on examples from the literature by a panel of experienced examiners. The rubric was used in the following year to assess theses and was evaluated thereafter. Based on this evaluation, adjustments were made that led to the rubric that is evaluated in this study (**Supplementary S1**). This rubric describes criteria for 13 categories typical for theses in the science domain (i.e. title, abstract, introduction, scope, structure, discussion, conclusions, scientific quality, spelling and grammar, style, length and lay-out, figures, references) and the writing process (professional attitude). Each category is further divided into 1-6 subcriteria (see Appendix 1). Examiners can tick boxes for each of these subcriteria to judge a subcriterion as insufficient, sufficient or good. It is allowed to tick boxes in two subcriteria, if the supervisor deems the level of the student is between insufficient/sufficient or between sufficient/good.

Students are graded from 1.0 (lowest) to 10.0 (highest). If the thesis is marked a 5.5 or higher the students receive 7.5 European Credits (ECs). If the examiners grade the thesis as insufficient (<5.5), the student need to improve the work and is allowed to hand-in a much-revised version again. If the examiners deem that it is unlikely that the required level is met by modifying the work, the student needs to start over with the thesis course and must find a different supervisor and examiner.

The rubric is purposely designed as an evaluation guideline for the examiners and as feedback tool for the student. The examiners assign one overall judgment (grade) based upon the rubric criteria. However, the criteria scientific quality and professional attitude are highlighted to advice examiners to particularly consider these two criteria when to decide upon the grade. Although explicitly not designed as a calculation table, if all criteria of the thesis are assessed at the 'sufficient' level, the suggested grade is a 7.

Quantitative assessment of rubric scores

To allow quantitative evaluation required for this study, we numerated all subcriteria. A 1 was assigned if the subcriterion was judged as insufficient, a 2 when sufficient and a 3 if deemed good. If two boxes for one subcriterion were ticked the average was calculated (i.e. 1.5 or 2.5). A 0 (null) was assigned when



▲**Figure 2** Trendline (blue) of the correlation between thesis grade and the mean score of the 13 subcriteria of the assessed rubric. Boxes behind the trendline indicate the median and variation in rubric score per grade. Shown are the boundaries of the second and third quartile of the data distribution. Black bars within the boxes indicate the median and whiskers the Q1 and Q4 values within 1.5 times the interquartile range. Individual dots are outliers beyond the Q1 and Q4 intervals.

▼**Table 1** Descriptive statistics for criteria scores of all rubrics used in the assessment of bachelor theses in biology.

Rubric criterium	**N[a]**	**Range**	**Mean**	**S.D.**	**r**	R^2_{adj} [b]
Title	296	1.0-3.0	2.27	0.48	0.48	0.23**
Abstract	294	1.0-3.0	2.29	0.56	0.59	0.34**
Introduction	296	1.0-3.0	2.43	0.52	0.61	0.37**
Scope	281	1.0-3.0	2.32	0.60	0.56	0.32**
Structure	296	1.0-3.0	2.39	0.50	0.76	0.58**
Discussion	291	1.0-3.0	2.31	0.49	0.72	0.52**
Scientific quality	295	1.0-3.0	2.44	0.44	0.79	0.62**
Spelling and Grammar	290	1.0-3.0	2.49	0.53	0.50	0.24**
Style	296	1.0-3.0	2.39	0.46	0.72	0.52**
Length and Layout	293	1.0-3.0	2.42	0.51	0.59	0.34**
Figures	296	1.0-3.0	2.36	0.53	0.62	0.38**
References	296	1.5-3.0	2.58	0.43	0.61	0.37**
Attitude	287	1.0-3.0	2.52	0.46	0.73	0.53**

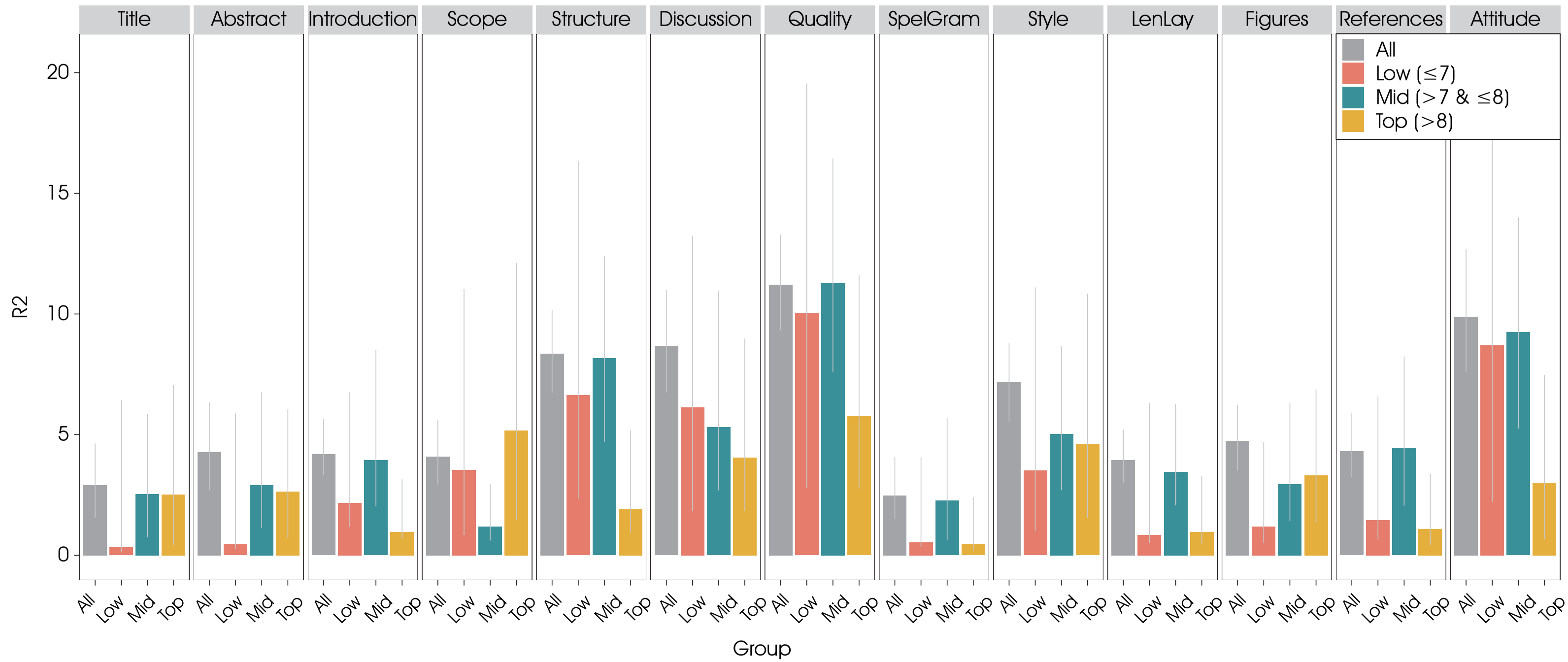
[a] The number of category scores differs per rubric category because of missing data
[b] * p<0.01 **p<0.001

5

none of the boxes of a particular category was ticked (i.e. left empty). The subcriteria within the introduction and scope category could not be directly numerated, since the number of tick boxes per assessment level differed; i.e. the insufficient category contains 2 boxes for introduction and 2 for scope, while respectively 4 and 2 subcriteria describe the introduction at the sufficient and good assessment levels, and 1 at each level the scope. To circumvent confounding effects, we therefore assigned a 1, 2, or 3 to the introduction and scope categories as a whole, without taking the subcriteria into account. A 1 was for instance assigned if only box(es) at the insufficient assessment level was/were ticked, a 2 if only boxes at the sufficient level was/were ticked and a 3 if only boxes at the good level was/were ticked. A 1.5 or 2.5 were assigned if boxes in more than one category was ticked, even if for instance 4 boxes of the introduction category were ticked at the sufficient assessment level and only one at the insufficient level. A similar method was applied to the title category and the conclusion part of the conclusions and discussion category, that both contained 2 subcriteria at the insufficient level and only 1 at the sufficient and good assessment levels. Besides the numeration of individual subcriteria, we calculated per rubric which (sub)categories were ticked at two assessment levels (i.e. between insufficient and sufficient or between sufficient and good. In addition, we calculated the average score per category.

Method of analysis

All criteria scores and grades were collected in a data matrix and analysed in R (http://www.R-project.org/), using the Openxlsx() function from the Openxlsx package (https://github.com/ycphs/openxlsx). Cronbach's alpha was determined using the alpha() function from the Psych package (Revelle 2019). Means and standard deviations (SDs) were calculated using base-R functions mean(), and sd(). For linear regression of criteria scores explaining the final grade, the summary(), and lm() functions of base-R were used. To estimate the relative importance of criteria scores, a linear model and the boot.relimp() and booteval.relimp() functions from the relaimpo package (Grömping 2006) were used. For visualization the packages: Ggplot2 (Wickham 2016), Cowplot (Wilke 2019) Hplots (Warnes et al. 2019), Hgdendro (de Vries and Ripley 2016) and Patchwork (Pedersen 2019) were used. The heatmap.2() function from



▲**Figure 3** Multiple linear regression analysis of thesis grade and criteria scores for all theses and for theses graded as sufficient (≤7.0), good (>7.0 and ≤8.0), and excellent (>8.0). The grey bars indicate standard deviations.

the Gplots package was used to calculate the dendrogram which was then extracted and used for visualisation using ggplot2 and patchwork.

Results

Descriptive statistics of criteria scores and thesis grades

The 318 bachelor theses in this study were assessed with an average grade of 7.58 (SD=0.80), ranging from 5.50 to 10.00. The rubrics used to guide assessment contained one or more subcriteria per criterion that could be scored with either a 1, 2 or 3 (insufficient, sufficient and good). The subcriterion scores were generally most similar to subcriterion scores belonging to the same criteria (**Figure 1**), suggesting that criteria stand alone and no splitting or merging of categories is needed for further analysis of the criteria scores. The criteria scores had high reliabilities for the final grade, Cronbach's α=0.96.

Regression of criteria scores for all thesis grades

To investigate the relation between rubric criteria scores and thesis grade, we plotted the average scores of all criteria (rubric score) against thesis grade (**Figure 2**). A positive relation is observed as expected, although the curve is sigmoidal. Nevertheless, rubric scores vary largely per specific thesis grade (**Figure 2**). A possible explanation for the variations is that some criteria contribute more than others. Hence, we determined how each of the separate criteria scores are related to thesis grade, while correcting for the interdependence of the criteria scores (**Table 1**). All criteria scores are significantly related to final grade. Since the scientific quality and professional attitude were marked in the rubric as "important", it was hypothesized that these criteria had highest effects on the thesis grade. The scientific quality and professional attitude are, together with structure, indeed the best predictors of grade and explain respectively 62%, 53% and 58% of the variance in grades (R^2_{adj} =0.62, R^2_{adj}=0.53 and R^2_{adj} =0.58). Title and spelling and grammar scores give the lowest prediction of grade and explain respectively 23% and 24% of the variance in grades. Multiple regression analysis of criteria on thesis grade

▼Table 2 Regression analysis of thesis grade bins and criteria scores adjusted for interdependence between criteria scores. The regression analysis is performed separately for three bins; sufficient thesis grades (left column, <7.0; n=48), good thesis grades (grades ≥7.0 to ≤8.0; n=158) and excellent thesis grades (grades > 8 ; n=55).

		Sufficient thesis grade bin (≤7.0)					Good thesis grade bin (>7.0 to ≤8.0)					Excellent thesis grade bin (>8.0)				
	Criterion	N[a]	S.D.	Mean	R	R^2_{adj} [b]	Na	S.D.	Mean	R	R^2_{adj} [b]	Na	S.D.	Mean	R	R^2_{adj} [b]
Opening	Title	93	0.29	2.00	0.10	0.00	144	0.47	2.28	0.21	0.04	59	0.48	2.64	0.17	0.01
	Abstract	93	0.38	1.89	0.14	0.01	142	0.52	2.36	0.29	0.08**	59	0.47	2.75	0.37	0.13*
	Introduction	93	0.43	2.03	0.25	0.05	144	0.46	2.53	0.39	0.15**	59	0.32	2.82	0.24	0.04
	Scope	85	0.54	1.89	0.33	0.10*	139	0.52	2.41	0.11	0.00	57	0.45	2.78	0.29	0.07*
Content	Structure	93	0.37	1.89	0.48	0.22**	144	0.38	2.53	0.48	0.23**	59	0.23	2.84	0.24	0.04
	Discussion	92	0.40	1.92	0.40	0.15**	141	0.39	2.36	0.54	0.29*	58	0.32	2.81	0.32	0.09
	Sc. Quality	92	0.34	2.00	0.55	0.30**	144	0.32	2.55	0.50	0.25**	59	0.22	2.87	0.38	0.13*
Presentation	Spel&Gram	91	0.50	2.16	0.24	0.05	140	0.49	2.58	0.27	0.07*	59	0.41	2.80	0.27	0.06
	Style	93	0.35	1.97	0.42	0.17**	144	0.36	2.48	0.39	0.15*	59	0.30	2.81	0.38	0.13*
	Length and Layout	91	0.44	2.03	0.22	0.04	144	0.45	2.51	0.35	0.12**	58	0.34	2.79	0.20	0.02
	Figures	93	0.42	1.99	0.33	0.10*	144	0.50	2.41	0.38	0.14**	59	0.30	2.82	0.27	0.06
	References	93	0.37	2.22	0.27	0.06*	144	0.37	2.69	0.28	0.07**	59	0.22	2.88	0.16	0.00
Attitude	Professional attitude	89	0.39	2.08	0.53	0.27**	141	0.34	2.64	0.39	0.15*	57	0.18	2.91	0.10	0.00

Note. N is total number of cases, S.D. is standard deviation, R^2adj is explained variance of thesis grade adjusted for the number of predictors
[a] The number of criteria scores differs per rubric category because of missing data.
[b] *p<.01 **p<.001

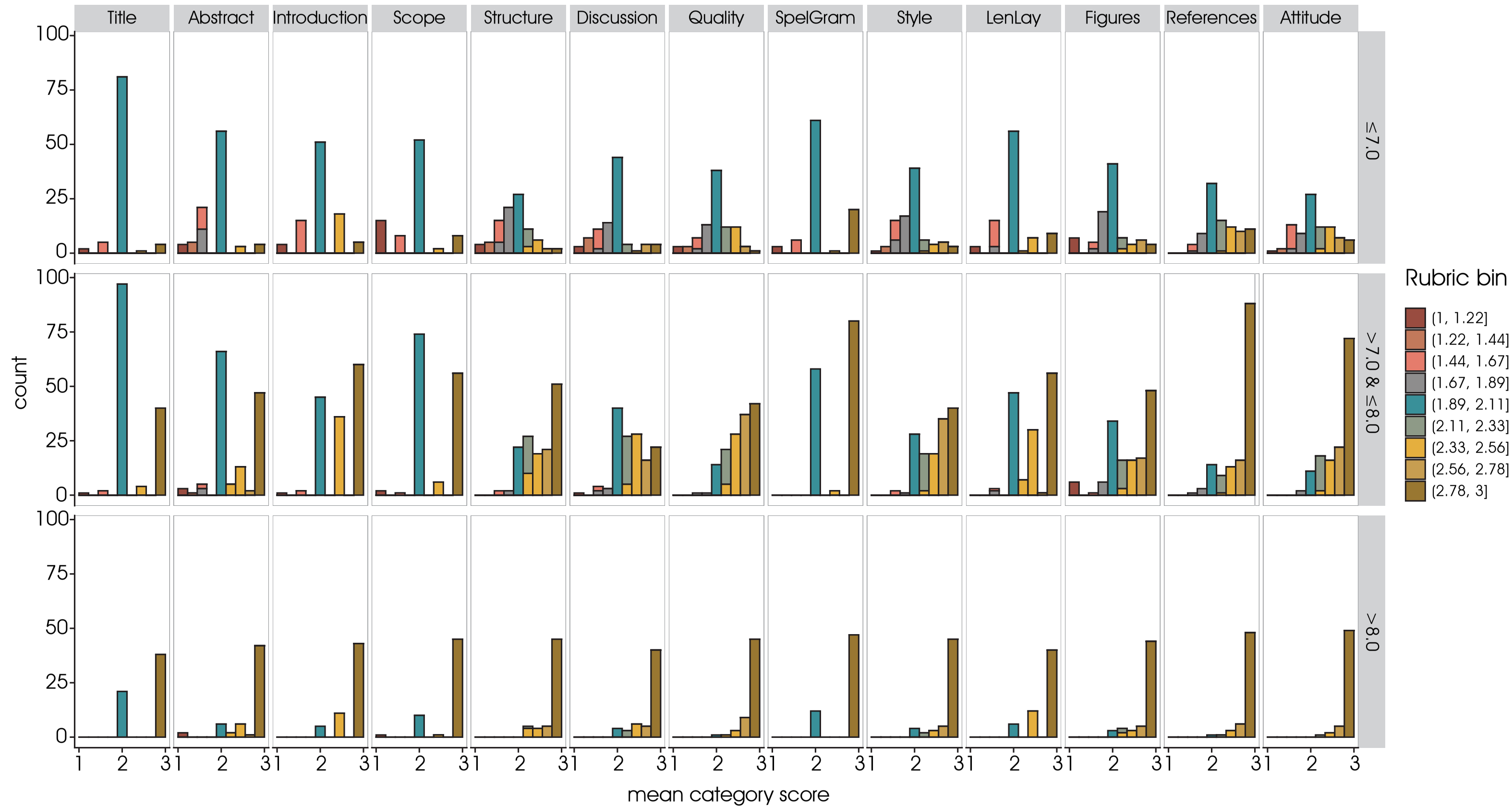
5

yielded similar results (**Figure 3**).

Regression of criteria scores for sufficient, good and excellent thesis grades

Although informative, overall correlations provide limited insight into which categories contribute most to (just) sufficient, good and excellent grades. We therefore assessed whether some rubric categories are of lower or higher relevance to examiners when assessing the student with a low or high grade. Therefore, we divided all theses in three groups (bins) based on the grade: theses with grades up to and including 7.0 (sufficient), above 7.0 and up to and including 8.0 (good) and above 8.0 (excellent). We then again performed a regression analysis on all categories, but now for these three bins separately (**Table 2**).

The regression analysis shows that scientific quality ranks among the best predictors for thesis grades in all three bins. Nonetheless, some categories are better predictors for grades in the sufficient bin than for those in the good or excellent bins. For example, theses in the sufficient bin are best distinguished by scientific quality (R^2_{adj} = 0.30) and professional attitude (R^2_{adj} = 0.27) scores. However, theses that received grades in the good bin are mainly distinguished by scientific quality (R^2_{adj} = 0.25) and discussion (R^2_{adj} = 0.29), whereas theses in the excellent bin are mainly distinguished by scientific quality (R^2_{adj} = 0.13), abstract (R^2_{adj}=0.13) and style (R^2_{adj} =0.13). Interestingly, abstract is one of the most predictive criteria in the excellent bin (R^2_{adj}=0.13), but one of the least predictive criteria of the sufficient bin (R^2_{adj}=0.01). Similarly, professional attitude is one of the most significant criteria (R^2_{adj}=0.27) for the sufficient bin, although being least significant for grades belonging to the excellent bin (R^2_{adj}=0.00). Thus, the regression analysis shows that some criteria appear more relevant in the assessment of thesis judged as sufficient than those that received an assessment that qualifies as good or excellent.



▲**Figure 4** Frequency distribution of criteria scores of the 13 rubric categories for theses graded as sufficient (≤7.0), good (<7.0 and ≤8.0), and excellent (>8.0). Criteria scores are color coded from red (criteria scores from 1 to 1.22) to green (criteria scores from 2.78 to 3).

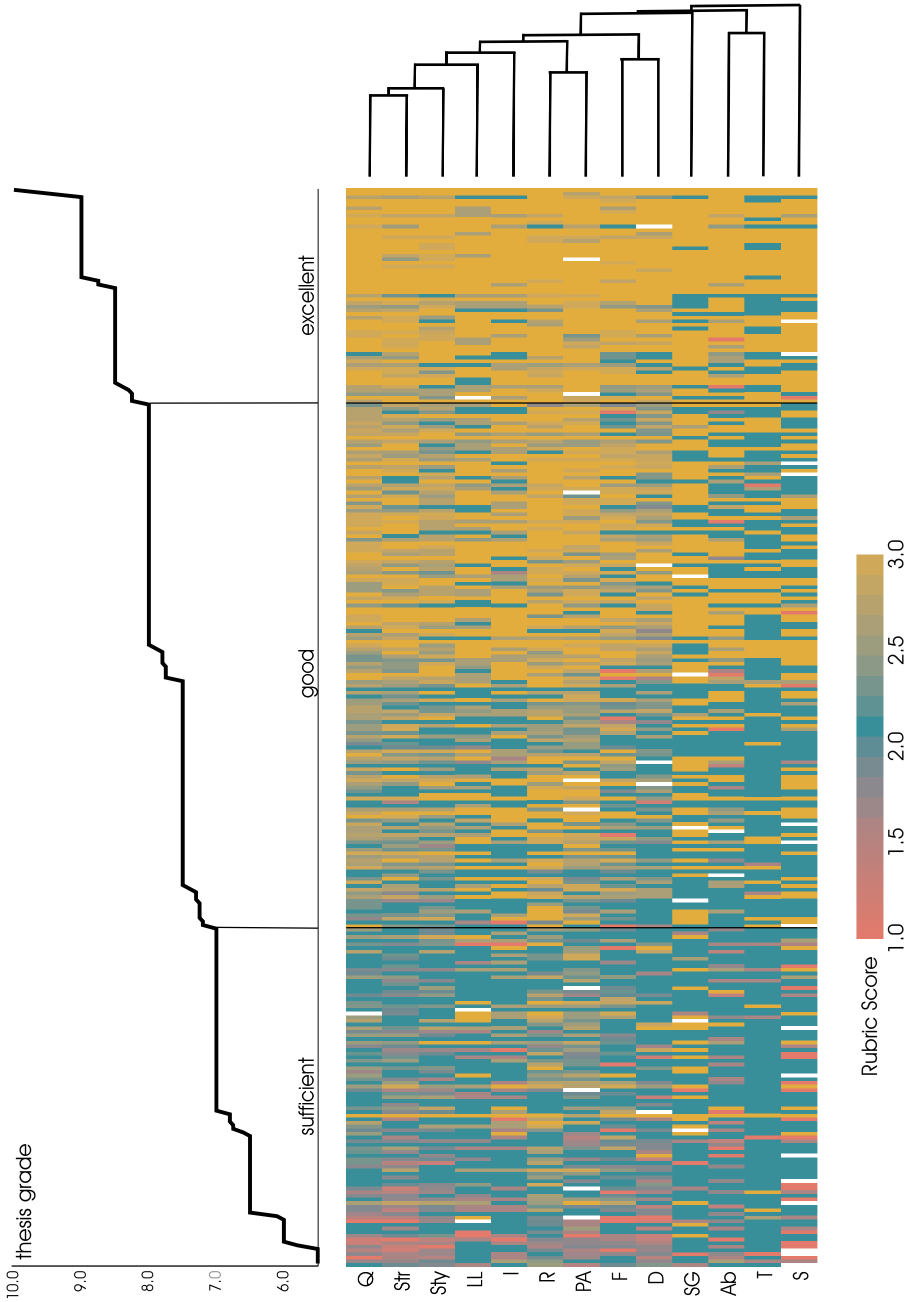
5

Visualization of criteria scores

The differences in the explained variances of criteria scores per grade bin suggest that some criteria appear of significance when assessing a sufficient, good or excellent thesis, or that some criteria are considered more determinative for an examiner to assign a grade that falls in a particular bin. We hence visualized the criteria scores per bin (**Figure 4**), and of all theses separately (**Figure 5**), to better understand the criteria scores provided. In Figure 4, the number of theses with a specific criteria score are shown for every bin. In Figure 5, all 318 theses are sorted in columns from the lowest thesis grade (left) to the highest grade (right). Rows represent colour-coded criteria scores with red bars for scores of 1.0, blue bars for scores of 2.0 and yellow bars for scores of 3.0. For example, the thesis with the lowest thesis grade (first column on the left) was scored with a 1.43 for professional attitude (red/blue) and a 2.40 for references (blue/yellow). The thesis with the highest thesis grade (last column on the right) was scored with a 3.00 for every single criterium. Figure 5 visualizes that students can obtain nearly insufficient thesis grades even if their use of references is scored above 2.0 on average (i.e. yellows bars are visible on the far-left side of the figure). Additionally, Figure 4 and 5 both reveal that scores above 2.0 for title, abstract, scope and/or spelling and grammar are apparently not required to obtain excellent thesis grades (i.e. red/blue bars are visible in the excellent bin of Figure 4 and on the far-right side of Figure 5). However, students never receive excellent thesis grades if their professional attitude and structure are scored with a 2.0 or below (i.e. only yellow bars are visible in the excellent bin of Figure 4 and on the far-right side of Figure 5).

Discussion

Bachelor theses are essential for the completion of many study programmes and it is logically relevant that assigned grades reflect the quality of the students' work and attitude in the best possible manner. Our quantitative analysis provides insight in the assessment of bachelor theses with the aid of rubrics. We demonstrate the significance of each rubric criterion on the overall thesis grades, and separately on the grades of thesis deemed sufficient, good or excellent.



▲**Figure 5** Average rubric criteria scores (rows) of all 318 assessed biology bachelor theses (columns). The 318 bachelor theses are sorted in columns from lowest thesis grade (left) to highest thesis grade (right) with black vertical lines marking the boundaries of the three separate grade bins: sufficient (≤7.0), good (<7.0 and ≤8.0), and excellent (>8.0). The criteria scores of each thesis are presented with color codes (red=sufficient, blue=good and yellow=excellent). Criteria scores that were missing are marked in white. The criteria are ordered based on the similarity in scores, as indicated by the dendogram on the right. Criteria that are scored similarly, such as quality and structure or references and attitude, are directly connected in this dendogram. The dendogram is calculated based on Euclidean distance hierarchically clustering.

5

Theses that are assessed with the same grade can still vary in their average rubric scores and vice versa; identical rubric scores can result in different grades, depending on the relative contribution the examiners assign to different criteria. One explanation of the variance in total rubric scores is that some of the rubric categories appear more significant in the assessment than others. Indeed, we demonstrate that e.g. scientific quality and structure are better predictors for thesis grade than other criteria. The high predictability of scientific quality for assigned grade was expected since, as 'rule of thumb', examiners are encouraged to put relatively more weight in their assessment to this criterion (**Supplementary S1**), together with professional attitude. Unfortunately, we cannot assess to what extent this encouragement contributed to the high level of predictability of scientific quality for thesis grade. The result is nonetheless in agreement with a study by Lundgren et al. (2008), who show that examiners mainly assess on students' ability to compare results of different studies; one of the subcriteria of scientific quality in the rubric of the current study. The other subcriteria defined for scientific quality are 2) demonstration of a high level of knowledge, 3) adequate discussion of the results, 4) well explained and supported arguments, and 5) clear distinction of facts and hypotheses. It would not be suprising if these subcriteria on scientific reasoning would also be the highest predictors for thesis grade if examiners were not informed on its importance. Most examiners in this study are namely tenured scientists and expected to assess theses from a researcher's perspective with focus on the students' line of reasoning. After all, the experience of the examiner appears an important factor for assessments (Kapborg and Berterö 2002; van der Vleuten et al. 2012).

This study also reveals that some criteria are worse or better predictors for sufficient grades than for good or excellent grades, suggesting prioritization in the examiner's mind. Interestingly, professional attitude strongly predicts grades of sufficient theses. Bachelor theses also were never deemed excellent if a student scored low for professional attitude, suggesting once more that attitude is an important criterion for thesis assessment. It should be mentioned that the professional attitude includes subcriteria on 1) required feedback, 2) response to feedback, 3) enthusiasm and commitment, 4) fulfilling agreements and 5) fulfilling deadlines. Thus, we propose that if the content of the thesis is only just sufficient, but the student is independent, committed, enthusiastic, adjusts feedback and fulfils agreements, examiners are prone to lift the grade from sufficient to good. Of note, the subcriteria of professional attitude on commitment, agreements and deadlines are independent of the quality of the thesis product itself and one can debate whether these manners should be considered when assessing theses. However, the first subcriteria on required feedback discusses whether the thesis is a true product of the student itself and this level of 'self-regulatory ownership' is regarded essential for students in higher education (Prins et al. 2017).

Remarkably, we found that the abstract is a better predictor for excellent grades than for sufficient or good grades. We hypothesize that, because the abstract is the first text the examiner reads, the quality of the abstract provides the first impression of the work, that may set the tone for the rest of the assessment, consciously or unconsciously. Alternatively, as the abstract is often the finishing touch of a thesis, it is only considered by the examiner in the final product, if the rest of the thesis and prior drafts already appeared to be excellent.

Nonetheless, we note that all criteria scores are relatively less predictive for the assessment of excellent theses compared to theses with sufficient or good grades. This can be explained by the distribution of rubric scores, showing that most criteria of excellent theses are assessed with the highest possible score of 3.0 (**Figure 4**). Thus, no further differentiation in rubric score can be made between criteria that only 'just fulfilled' the description of the highest rubric category or criteria that 'outreach' this description by far. Moreover, as the total number of cases is relatively low in the excellent grade bin, distinguishing of rubric scores between 'just excellent grade' theses scored with an 8.0 and with 'highly excellent theses' scored with a 9.0 or 10.0 is also hampered by restrictions in statistical power. Alternatively, factors that are not included in the rubric play a role in the assessment.

Assessment procedure

The significance of the rubric criteria provides insight into the criteria that examiners deem relevant. However, this study does not disclose the implicit decisions that examiners make while assessing. We therefore encourage additional studies wherein examiners explain if and how they assign thesis grades with the aid of rubrics. This might reveal whether additional implicit factors (not described in the rubric) are considered during assessment and might explain if and how the rubric is used when deciding upon on overall grade. For example, examiners may fully rely on the rubric to determine the thesis grade, but it might also be that examiners simply use rubrics to justify the grade they already have in mind (Bloxham et al. 2011; Timmerman et al. 2011). Most likely, our dataset covers the whole spectrum in between these extremes. It is naturally expected that examiners have different assessment approaches, and this might be particularly true for the biology theses of the current study that are assessed by examiners with diverse experiences in education, supervision and assessment and moreover, are from diverse disciplines with each their own habits and cultures. Differences in final thesis grades have already been confirmed for external and internal examiners (Williams and Kemp 2019), for examiners with low and high affiliation with the student (de Kleijn et al. 2012) and for examiners that did or did not supervise the student themselves (Lundgren et al. 2008). Besides, it appears that during assessment, inexperienced examiners are more focused on each separate institutional criterion and that experienced examiners consider the interrelatedness of the criteria more often (Adnan and Bulut 2014; Kiley and Mullins 2004; Sadler 2009). It would be interesting to study whether the rubric criteria scores are also different between these examiners and whether experienced examiners give more or less weight to certain criteria in comparison with inexperienced examiners. We should however stress that we do not recommend using a fixed formula to calculate a final theses grade from criteria scores. We encourage to leave the overall judgment of theses to the supervisors since i) they are the experts, ii) they can assess the overall quality and consider how criteria congregate and iii) they can judge how relevant each criterion is in their specific discipline. Nonetheless, in order to assess theses more similarly, we encourage moderation sessions wherein examiners explain each other their understanding of the rubric criteria and discuss how they decide upon the final grade (Grainger et al. 2016).

Practical implications

The current study suggests that all criteria matter, showing that students should pay attention to every criterion. However, the regression analysis suggests that students should prioritize working on scientific quality, structure and professional attitude above criteria such as a correct use of spelling and grammar and a catchy title that justifies its content. Thus, students are advised to be enthusiastic and committed (professional attitude) and explain, discuss and integrate the results (scientific quality) in a clear, consistent and structured manner (structure). Supervisors are similarly advised to put emphasis on the most significant criteria and give extra feedback and guidelines to students on the structure and scientific quality of a thesis. Likewise, supervisors can improve the overall thesis quality of their students by giving extra guidance on criteria that are hardly scored as excellent such as title, abstract and scope.

Regression analyses such as performed in this study are of specific interest for boards of examiners and education managers alike, as they can help to assess whether the end terms of their curriculum are achieved and whether the criteria they find of highest importance, also are most predictive for assigned grades. Fortunately, scientific quality is thought to be the most important category according to the rubric developers and also appears the best predictor for thesis grade. Similarly, professional attitude is highlighted as important criterion and appears to be highly predictive for sufficient thesis grades as well. We do however not know if and how the highlights in the rubric assessed in this study has contributed to the assessments. Either way, these results suggest that examiners primarily assess on the criteria that are also found to be most relevant by the board of examiners and institutional directives. We thus recommend performing similar analyses from time-to-time to estimate whether the assessment reflects the quality of the students' work and validate if assignments are mainly assessed on the criteria that are considered to be most relevant.

Conclusions

This explorative study presents a quantitative analysis from a large dataset of rubric scores and the corresponding thesis grades. The difficulty of assigning thesis grades is that many skills need to be considered to provide one single overall grade that fits all. We show that all rubric criteria scores are predictive for thesis grades, suggesting that all criteria are relevant for assessment. Furthermore, thesis grades are best distinguished by scientific quality and structure. In addition, the distinct analysis of criteria scores for separate grades showed that some criteria appear more predictive for low grades than for high grades. Information on these criteria might support students to prioritize on the most significant criteria, guide supervisors to give their students extra advice on these criteria and help education managers and accreditation committees to evaluate if the assessments are in agreement with the criteria they find of highest relevance.

Informed consent and statement of human rights

All procedures performed in this research, involving human participants, were in accordance with the ethical standards of the institutional and/or national research committee (Ethics Review Board of Beta and Geosciences at Utrecht University, reference number Bèta D-2031; available on motivated and reasonable request).

Supplemental figures and tables

▼**Supplementary S1** Rubric used for assessing bachelor biology theses at Utrecht University

Criteria	Insufficient	Sufficient	Good
Title	T1 ▶ Title does not justify the content T1 ▶ Title is too long	T1 ▶ Title justifies the content	T1 ▶ Title justifies the content and is catching
Abstract	Ab1 ▶ Abstract is missing Ab2 ▶ Abstract is unclear Ab2 ▶ Abstract lacks, or gives too much, irrelevant information	Ab1 ▶ The abstract is present Ab2 ▶ In general the abstract is clear Ab2 ▶ The abstract contains all important elements	Ab1 ▶ Abstract is present Ab2 ▶ Abstract contains a clear, concise and balanced description of all important elements
Introduction and scope of thesis	I1 ▶ The introduction contains incomplete information I2 ▶ The introduction contains a lot of irrelevant information I3 ▶ The scope of the thesis is missing I3 ▶ The scope of the thesis is unclear	I1 ▶ The introduction contains sufficient information I1 ▶ The introduction is functional I2 ▶ Most of the presented information is relevant I2 ▶ Some repetition is present, but this is not disturbing I3 ▶ The scope of the thesis is clear	I1 ▶ The introduction contains sufficient relevant, information to put the scope in a broader perspective I2 ▶ The introduction is catchy and inviting to continue reading I3 ▶ The scope of the thesis is well defined

▼

5

Criteria	Insufficient	Sufficient	Good
Structure	Str1 ▶The order of paragraphs is illogical. Str2 ▶ Transitions between the paragraphs are lacking Str3 ▶ The different sections are not in balance Str4 ▶ Not all of the information is consistent with the scope Str5 ▶ Information is frequently repeated Str6 ▶ The line of thought is difficult to follow	Str1 ▶ The structure and order of most paragraphs is sound Str2 ▶ Transitions between paragraphs are mostly logical Str3 ▶ Most sections are well balanced Str4 ▶ Most of the information is consistent with the scope Str5 ▶ Information is sometimes unnecessarily repeated Str6 ▶ The line of thought is clear	Str1 ▶ The structure and order of the paragraphs is sound Str2 ▶ Smooth transitions between paragraphs Str3 ▶ The sections are well balanced Str4 ▶ The information is consistent with the scope Str5 ▶Only if necessary, the information is repeated Str6 ▶ The line of thought is clear and well structured
Discussion / Conclusion	D1 ▶The discussion is missing or too short D2▶ The discussion lacks depth D3 ▶ The discussion does not go back to the scope of the thesis D4 ▶ No conclusions are drawn D4 ▶The conclusions are not supported by the presented findings	D1 ▶ The discussion ties together loose ends D2 ▶ The discussion has limited depth D3 ▶The discussion goes back to the scope of the thesis D4 ▶ Conclusions are drawn in line with the presented findings	D1 ▶ Discussion demonstrates a good overview of the subject of the thesis D2 ▶In depth discussion D3 ▶ The discussion goes back to the scope of the thesis D4 ▶ Conclusions are drawn in line with presented findings

▼

Crriteria	Insufficient	Sufficient	Good
Scientific quality	SQ1 ▶ The thesis does not demonstrate sufficient knowledge of the field SQ2 ▶ Results from the referred articles are not discussed correctly SQ3 ▶ Arguments are not supported by evidence SQ4 ▶ It is difficult to distinguish facts from hypotheses SQ5 ▶ Information is not well integrated: the text often reads as loose pieces of information	SQ1 ▶ The thesis shows sufficient knowledge of the field in most parts SQ2 ▶ Most important results from the references are discussed correctly SQ3 ▶ Some arguments are supported by evidence SQ4 ▶ Facts and hypotheses are not always easy to distinguish from each other SQ5 ▶ Limited integration of information	SQ1 ▶ The thesis shows a high level of knowledge and understanding of the field SQ2 ▶ Results from the references are adequately discussed SQ3 ▶ Arguments are well explained and scientifically supported SQ4 ▶ Facts and hypotheses can easily be distinguished SQ5 ▶ Results from different papers are integrated well into new insights
Spelling and grammar	SG1 ▶ Disturbing spelling or grammar mistakes, which complicates understanding of the text	SG1 ▶ There are errors in spelling and/or grammar, but this does not prevent understanding the text.	SG1 ▶ None, or very few errors in grammar or spelling
Style	Sty1 ▶ Writing style is hardly scientific Sty2 ▶ Many sentences are not structured properly, or too long or too short Sty3 ▶ Little variation in vocabulary Sty4 ▶ Important terms are not explained Sty5 ▶ Bad use of punctuation	Sty1 ▶ In general the writing style is scientific Sty2 ▶ Few sentences are badly structured, but this is not too disturbing Sty3 ▶ In general a sufficient vocabulary is used Sty4 ▶ Important terms are explained Sty5 ▶ Use of punctuation is usually correct and makes the text easier to follow	Sty1 ▶ Style is concise, and scientific. It is a pleasure to read the text Sty2 ▶ Sentences flow smoothly and are not too long or too short Sty3 ▶ Rich vocabulary Sty4 ▶ Important terms are well explained Sty5 ▶ Good use of punctuation

▼

Criteria	Insufficient	Sufficient	Good
Length /Lay-out	LL1 ▶ The length is too short / too long LL2 ▶ The layout is sloppy and not appealing	LL1 ▶ The length is as required LL2 ▶ The layout is uniform and organized	LL1 ▶ The length is as required LL2 ▶ The layout is uniform and well organized
Figures	F1 ▶ Insufficient number of figures F2 ▶ Figures are often of bad quality F3 ▶ Too many irrelevant figures F4 ▶ Figure legends are missing or badly formulated F5 ▶ The figures are not referred to in the text or the reference is wrong F6 ▶ A clear explanation of the figure in the text is missing	F1 ▶ Sufficient number of figures F2 ▶ Most figures are of good quality F3 ▶ Most figures are relevant F4 ▶ Figure legends are adequate to understand the figure F5 ▶ Good reference to the figures in the text F6 ▶ Most figures are explained well in the text	F1 ▶ Sufficient number of figures, that support the text F2 ▶ Figures are of good quality F3 ▶ Figures are relevant and when needed adjusted or specifically designed for the thesis F4 ▶ Figure legends are complete F5 ▶ Good reference to the figures in the text F6 ▶ Figures are explained well in the text
References	R1 ▶ Insufficient number of references (<20) R2 ▶ An insufficient number of primary literature is used R3 ▶ Finding and selecting references was mainly done by the supervisor R4 ▶ None, or incorrect referral to references in the text R5 ▶ The reference list is incomplete or sloppy	R1 ▶ Sufficient number of references (>20) R2 ▶ A sufficient amount of primary literature was used R3 ▶ Most sources were found by the student R4 ▶ Most of the articles are correctly referred to in the text R5 ▶ The reference list is largely complete, clear and uniform, with some minor mistakes	R1 ▶ Sufficient number of references (>20) R2 ▶ A sufficient amount of primary literature was used R3 ▶ All references were found by the student R4 ▶ The articles are correctly referred in the text R5 ▶ The reference list is complete, clear and uniform

▼

Criteria	Insufficient	Sufficient	Good
Professional attitude	PA1 ▶ A lot of supervision was required PA2 ▶ Minimal improvement after feedback PA3 ▶ The student did not put enough effort into the thesis work PA4 ▶ The student failed to follow the agreements PA5 ▶ The student failed to meet the deadlines	PA1 ▶ Regular feedback sessions were required PA2 ▶ Feedback led to reasonable improvements PA3 ▶ The student was committed PA4 ▶ The student fulfilled the agreements PA5 ▶ The students met the agreed deadline	PA1 ▶ Minimal feedback was required, the thesis is a true product of the student PA2 ▶ Response to feedback yielded excellent improvements PA3 ▶ The student was enthusiastic and committed PA4 ▶ The student fulfilled the agreements PA5 ▶ The student met the agreed deadline
Further clarifications			

The different check-boxes in a cell should be read as and/or.

This table is developed to help in evaluating the thesis and in giving feedback to the student. The examiners together decide the weight of the different components and how they translate into the final grade. Should all criteria of this thesis but judged "sufficient", than the suggested mark is a 7.

Two important criteria for the Bachelor degree (Scientific quality and professional attitude) are indicated in black.

6

Discussion

Discussion

The studies underlying this thesis aimed to improve our general understanding on learning and assessment related to current trends and questions in the field of biology higher education. Education is continuously changing as the context of learning develops, learning mechanisms are better understood, other topics become more relevant and new technical tools become available. As a result, there is a continuous need of research to evaluate the learning effects and mechanisms of these trends and further improve the present education.

Changing learning environment

An example of a current and global change of the learning environment is the increasing number of students in higher education (Schofer & Meyer 2005). This results in large student populations at some universities, which requires new lesson activities to keep large groups of students actively involved. Another development is that many universities currently promote both internationalization and interdisciplinarity (de Wit & Altbach 2020; Tobi & Kampen 2018). This development asks for more studies on how students can effectively learn to work with other disciplines and how teachers can promote learning for a diverse group of students with various sociocultural backgrounds.

A good example of a factor with a sudden and profound effect is the covid19 pandemic. The need for distance learning during the pandemic increased teachers' interests for teaching through video (Lowenthal et al. 2020). In addition, the measurements during the covid19 pandemic limited the amount of laboratory activities allowed at campus, which asked for the design of alternative activities (Pols 2020; Ray & Srivastava 2020). These experiences resulted in new insights on laboratory activities and their learning objectives. Teachers had to make choices and decide for which objectives the lab activities were necessary and for which objectives other activities were equally or even more sufficient. It is expected that these experiences will result in a revolution in laboratory education. In Chapter 4, we addressed the problem that expected learning outcomes in laboratory education are not always met. We argued that students are regularly required to perform many new and different tasks when doing experiments in the lab, such as recalling information, making observations, searching for materials, following procedures, interpreting data and maintaining a lab journal. These different tasks likely result in a cognitive overload and thus in fewer learning outcomes (Johnstone 1997). In order to increase learning outcomes, inquiry-based labs are currently making their return. The main idea of this type of laboratory education is that students design their own research set-up and thus have more time to think about the purpose of the experiment (Brownell & Kloser 2015). Inquiry-based labs are thus expected to improve understanding of the theoretical background of a lab activity and its relation to practice. Nonetheless, inquiry-based labs can be time-consuming and expensive and require that students already have some knowledge of lab techniques (Gormally et al. 2009; Johnstone & Al-Shuaili 2001; Krajcik et al. 1998; Wei & Woodin 2011). We showed that pre-lab activities improved students' understanding of the theory and its relation to practice during laboratory education and helped to correctly understand the results and draw conclusions. We hypothesize that these learning effects result from a lower cognitive overload, although other learning mechanisms might also be at play. Of course, lab teachers often have many learning objectives other than students' understanding of the theory such as designing an experiment, performing practical skills, keeping up a lab journal and discussing limitations of an experimental set-up. Current insights from lab education during the covid19 pandemic might stimulate teachers to reconsider their objectives and redesign laboratory activities such that they better fit the intended objectives.

New technical tools

The development of new technical tools provides new opportunities in teaching. An example is the increased use of knowledge clips in higher education, which resulted from advances in streaming technology (Kay, 2012). The use of knowledge clips has again led to an increase in the number of companies that enhances knowledge clips with additional interactive tools (HapYak 2021; Hihaho 2021; H5P 2021; Panopto 2021; PlayPosit 2021; Scalable Learning 2020). In other words, the use of interactive knowledge clips in higher education is a new trend that asks for more studies on their effects on learning. The studies in Chapter 2 and 3 of this thesis contribute to our current understanding of learning effects and mechanisms from one of the most popular interactive video tools: the possibility to insert questions within videos. These pop-up questions within videos were expected to improve retention as it is established that tests in general improve memory, known as the testing effect (Glover 1989; Karpicke & Blunt 2011; Pastötter & Bäuml 2014; Roediger & Butler 2011; Szpunar et al. 2008). Indeed, our studies show that the occurrence of pop-up questions promote students' test performance, which advocates the use of pop-up questions in general. In addition, we reveal how pop-up questions likely promote learning, namely by stimulating students to review and rewind the video material. We also showed that students report different learning mechanisms of pop-up questions when posted either before (pre-questions) and/or after (post-questions) the corresponding video fragment. In all conditions, most students claimed that pop-up questions increased their focus of attention, helped them to think about the content and test their understanding. However, some students that answered pre-questions additionally explained that pop-up questions activate their prior knowledge. In contrast, students that only answered these questions after the video fragment reported how pre-questions increased their focus towards related topics within the next video fragment. Although we only determined the perceived learning mechanisms, our results give an indication that increased attention is a very important learning mechanism. As traditional lectures are replaced with videos and the popularity of educational videos is still increasing, there is a need to find out how different students learn from video and how to optimize that learning process.

Note that chapter 2 and 3 of this thesis specifically focus on pop-up questions in voice-over presentations, but that many other factors of video are expected to affect learning. Video characteristics such as novelty, development in complexity and a rewarding closure are for example related to students' interest in the video, which is again related to their interest in its content (Wijnker 2021). Thus, factors such as video type and storyline are also expected to affect students' learning. Additional studies are recommended that indicate when and whether different learning mechanisms take place and how they affect learning from video.

Need for quality control

Besides continuous studies on current developments in education, there is also a need to continuously examine if education and assessment meet the objectives of the program. In Chapter 5 we provided an example of a quantitative evaluation on the assessment procedure of bachelor theses. This study showed that the final thesis grade had highest correlation with criteria scores for scientific quality and structure. The study also signifies that professional attitude was more predictive in the range of low grades, whereas the abstract was relatively more predictive in the range of high grades. We should note that the analysis in Chapter 5 does not show how teachers actually use the rubrics for the assessment of bachelor theses; teachers could for example use the rubric to determine a grade or to confirm a grade they already have in mind. Thus, rubrics alone do not guarantee correspondence in assessment across teachers. Nonetheless, Chapter 5 shows that the criteria scores that predict the final grade best are also the criteria that were considered by the faculty as being most significant for a good thesis. To conclude, there will always be a need to monitor the quality of assessment and therefore I would recommend study programs in all fields to regularly perform analyses such as in Chapter 5 to test whether the assessments are in alignment with the main objectives of that particular assignment.

Practical implications and future directions

The examples above show that education is continuously changing and that these developments asks for continuous research on its learning effects and mechanisms. From Chapter 2 and 3 we recommend teachers to include pop-up questions in their knowledge clips, since only the presence of these questions already promotes the retention of the video content. However, Chapter 3 shows that both pop-up questions before and/or after the explanation of that topic are equally effective, although students prefer the latter. From Chapter 4 we would recommend teachers to do a pre-lab activity that includes information on the background theory and on the set-up of the experiment in order to improve the learning outcomes of the laboratory activities. In Chapter 5 we show that statistical analysis of rubric data can give insight in the proper use of the rubric during assessment, and we want to stimulate teachers as well as education managers to use this type of analysis to confirm the quality of their assessment rubrics.

I do not only want to formulate a recommendation for practitioners but also researchers in the field of Biology Education Research (BER). Previous studies in biology education are mostly descriptive and mainly show whether something works (DeHaan 2011). We now need to shift to study how and why these things actually work. This recommendation is also given by other biology education researchers in the field and even got a name; BER 2.0 (Dolan 2015). I would like to promote this second generation of biology education research and call for more in-depth studies grounded on theories of learning and studying learning mechanisms of a diverse range of learning activities. If we better understand how learning mechanisms work, it will also be easier to design or adapt effective methods of teaching. In this thesis, I already made a shift towards the second generation of BER by exploring the underlying learning mechanisms of pop-up questions and discussing how cognitive overload might affect students' learning abilities in laboratory education.

In conclusion, this thesis provides new insights that can be used to improve teaching in general, especially in the field of biology higher education. During the research described in this thesis it became clear how difficult it can be for educational researchers to understand the needs, limits and the context that teachers experience. On the other hand, most teachers have limited time and knowledge to find research results that are relevant for them and translate these results into improvements for their teaching. Indeed, there seems to be a clear cut between events designed to share research on education and events designed to inform teachers (National Research Council 2012, p.37). Educational researchers mainly aim to publish in high quality research journals, whereas teachers rather read, if any, practitioner journals instead. Nonetheless, the demand for evidence-based teaching, the current popularity of the SoTL movement and the emergence of hybrid journals and conferences show that steps are undertaken to bring research and practice together. In addition, I believe that BER researchers that perform experiments within the field and are involved in biology education have a perfect expertise to make the translation between research and practice in the field of biology education. This research project enabled me to inspire fellow colleagues with ideas obtained from the educational literature and to update both supporting personnel and fellow researchers with current issues and questions of teachers. This experience leads me to my final recommendation; for departments I would recommend having such a translator that can provide evidence-informed information and practical tips based on the specific needs of their teachers. In the near future I hope to keep on inspiring both my students and my fellow colleagues. In any case, with the upcoming trends and current issues in biology higher education there are many topics left to investigate and communicate about.

&

Addendum

References

Adnan, K., and Bulut, O. (2014). Crossed random-effect modeling: Examining the effects of teacher experience and rubric use in performance assessments. *Eurasian Journal of Educational Research*, 57, 1–28. https://doi.org/10.14689/ejer.2014.57.4

Alba, J. W., and Hasher, L. (1983). Is memory schematic? - ProQuest. *Psychological Bulletin*, 93(2), 203–231. https://doi.org/10.1037/0033-2909.93.2.203

Anderson, C. W., Sheldon, T. H., & Dubay, J. (1990). The effects of instruction on college nonmajors' conceptions of respiration and photosynthesis. *Journal of Research in Science Teaching* (Vol. 27, pp. 761–776). http://www.ncbi.nlm.nih.gov/pubmed/19703223%5Cnhttp://doi.wiley.com/10.1111/j.1558-5646.2009.00825.x

Andrade, H., and Du, Y. (2005). Student perspectives on rubric-referenced assessment. *Practical Assessment, Research and Evaluation*, 10(3). http://books.google.ca/books?id=hkjOs2HtwB4C&printsec=frontcover&dq=intitle:Student+perspectives+of+assessment+Considering+what+assessment+means+to+learners+Gavin+Brown+Academia+edu&hl=&cd=1&source=gbs_api

Antony, J. W., Ferreira, C. S., Norman, K. A., and Wimber, M. (2017). Retrieval as a Fast Route to Memory Consolidation. *Trends in Cognitive Sciences*, 21(8), 573–576. https://doi.org/10.1016/j.tics.2017.05.001

Armstrong, H. (1898). On the heuristic method. Special Reports on Educational Subjects, 3(Eyre & Spottiswood).

Baepler, P., Walker, J. D., Driessen, M. (2014). It's not about seat time: Blending, flipping, and efficiency in active learning classrooms. *Computers and Education*, https://doi.org/10.1016/j.compedu.2014.06.006

Barral, A. M., Ardi-Pastores, V. C., Simmons, R. E. (2018). Student learning in an accelerated introductory biology course is significantly enhanced by a flipped-learning environment. *CBE—Life Sciences Education*, https://doi.org/10.1187/cbe.17-07-0129

Bishop, J. L., Verleger, M. A. (2013). The flipped classroom : A survey of the research. Resource Document. *American Society for Engineering Education*. http://www.asee.org/public/conferences/20/papers/6219/view

Bloom, B. S., Engelehard, M. D., Furst, E. J., Hill, W. H., Krathwohl, D. R. (1956). *Taxonomy of educational objectives: The classification of educational goals* (Handbook I). New York: David McKay Company.

Bloxham, S., Boyd, P., and Orr, S. (2011). Mark my words: The role of assessment criteria in UK higher education grading practices. *Studies in Higher Education*, 36(6), 655–670. https://doi.org/10.1080/03075071003777716

Bonde, M. T., Makransky, G., Wandall, J., Larsen, M. V, Morsing, M., Jarmer, H., & Sommer, M. O. A. (2014). Improving biotech education through gamified laboratory simulations. *Nature Biotechnology*, 32(7), 694–697. https://doi.org/10.1038/nbt.2955

Bourke, S., and Holbrook, A. P. (2013). Examining PhD and research masters theses. *Assessment and Evaluation in Higher Education*, 38(4), 407–416. https://doi.org/10.1080/02602938.2011.638738

Bouwmeester, R. A. M., Kleijn, R. A. M., ten Cate, O. T. J., Rijen, H. V. M., and Westerveld, H. E. (2016). How do medical students prepare for flipped classrooms? *Medical Science Educator*, 26, 53–60. https://doi.org/10.1007/s40670-015-0184-9

Brame, C. J. (2016). Effective educational videos: Principles and guidelines for maximizing student learning from video content. *CBE-Life Sciences Education*, https://doi.org/10.1187/CBE.16-03-0125

Britton, J. N., Martin, N. C., & Rosen, H. (1966). *Multiple markings of english compositions*. Evans.

Brownell, S. E., and Kloser, M. J. (2015). Toward a conceptual framework for measuring the effectiveness of course-based undergraduate research experiences in undergraduate biology. *Studies in Higher Education*, 40(3), 525–544. https://doi.org/10.1080/03075079.2015.1004234

Bull, S. G., and Dizney, H. F. (1973). Epistemic-curiosity-arousing prequestions: Their effect on long-term retention. *Journal of Educational Psychology*, 65(1), 45–49.

Butler, A. C. (2010). Repeated testing produces superior transfer of learning relative to repeated studying. *Journal of Experimental Psychology: Learning Memory and Cognition*, 36(5), https://doi.org/10.1037/a0019902

Campbell, N. A., Reece, J. B., Urry, L. A., Cain, M. L., Wasserman, S. A., Minorsky, P. V., Jackson, R. B. (2015). *Biology, A global approach* (10th Inter). Harlow: Pearson Education.

Campbell, N. A., Urry, L. A., Cain, M. L., Wasserman, S. A., Minorsky, P. V., Reece, J. B. (2017). *Biology, A global approach* (11th Inter). Harlow: Pearson Education.

Carpenter, S. K., Pashler, H., Wixted, J. T., and Vul, E. (2008). The effects of tests on learning and forgetting. *Memory and Cognition*, 36(2), 438–448. https://doi.org/10.3758/MC.36.2.438

Carpenter, S. K., and Toftness, A. R. (2017). The effect of prequestions on learning from video presentations. *Journal of Applied Research in Memory and Cognition*, 6(1), 104–109. https://doi.org/10.1016/j.jarmac.2016.07.014

Chan, J. C. K., McDermott, K. B., Roediger, H. L. (2006). Retrieval-induced facilitation: Initially nontested material can benefit from prior testing of related material. *Journal of Experimental Psychology: General*, https://doi.org/10.1037/0096-3445.135.4.553

Creswell, J.W. (2007). *Qualitative inquiry and research design: Choosing among five approaches* (Second edition). Sage Publications. https://doi.org/10.1177/1524839915580941

Cummins, S., Beresford, A. R., Rice, A. (2016). Investigating engagement with in-video quiz questions in a programming course. *IEEE Transactions on Learning Technologies*, 9(1), 57–66.

D'Avanzo, C. (2008). Biology concept inventories: Overview, status, and next steps. *BioScience*, 58(11), 1–7. https://doi.org/10.1641/B581111

Dawson, P. (2017). Assessment rubrics: towards clearer and more replicable design, research and practice. *Assessment and Evaluation in Higher Education*, 42(3), 347–360. https://doi.org/10.1080/02602938.2015.1111294

de Jong, T., Linn, M. C., & Zaccharia, Z. C. (2013). Physical and virtual Laboraties in science and engineering education. *Science*, 340(April), 305–308.

de Kleijn, R. A. M., Mainhard, M. T., Meijer, P. C., Pilot, A., and Brekelmans, M. (2012). Master's thesis supervision: Relations between perceptions of the supervisor-student relationship, final grade, perceived supervisor contribution to learning and student satisfaction. *Studies in Higher Education*, 37(8), 925–939. https://doi.org/10.1080/03075079.2011.556717

DeBoer, G. (2019a). A history of ideas in science education: implications for practice. Teachers College Press.

DeBoer, G. (2019b). *School science seeks its own identity. In A history of ideas in science education: implications for practice* (pp. 175–205). Teachers College Press.

Derry, S. J., Sherin, M. G., and Sherin, B. L. (2014). Multimedia learning with video. In R. Mayer (Ed.), *The Cambridge Handbook of Multimedia Learning, Second Edition* (pp. 785–812). Cambridge University Press. https://doi.org/10.1017/CBO9781139547369.038

DeHaan, R. L. (2011). Education research in the biological sciences: A nine decade review. Presented at the *Second Committee Meeting on the Status, Contributions, and Future Directions of Discipline-Based Education Research*, 1–24. http://www7.nationalacademies.org/bose/DBER_DeHaan_October_Paper.pdf

de Vries, A., and Ripley, B. D. (2016). *ggdendro: Create dendrograms and tree diagrams using "ggplot2"* (0.1-20). https://github.com/andrie/ggdendro

de Wit, H., & Altbach, P. G. (2020). Internationalization in higher education: global trends and recommendations for its future. *Policy Reviews in Higher Education*, 1–19. https://doi.org/10.1080/23322969.2020.1820898

&

Diederich, P. B., French, J. W., & Carlton, S. T. (1961). Factors in judgments of writing ability. *ETS Research Bulletin Series*, 1961(2), i–93. https://doi.org/10.1002/j.2333-8504.1961.tb00286.x

Dirks, C. (2011). The Current Status and Future Direction of Biology Education Research. *National Research Council Commissioned Paper*, January, 1–41. http://onlinelibrary.wiley.com/doi/10.1002/cbdv.200490137/abstract%5Cnhttp://www7.nationalacademies.org/bose/1DBER_Dirks_October_Paper.pdf

Dolan, E. L. (2015). Biology education research 2.0. *CBE Life Sciences Education*, 14(4), 1–2. https://doi.org/10.1187/cbe.15-11-0229

Domin, D. S. (1999). A review of laboratory instruction styles. *Journal of Chemical Education*, 76(4), 543–547. https://doi.org/10.1021/ed076p543

Domingues, L., Rocha, I., Dourado, F., Alves, M., & Ferreira, E. C. (2010). Virtual laboratories in (bio)chemical engineering education. *Education for Chemical Engineers*, 5(2), 22–27. https://doi.org/10.1016/j.ece.2010.02.001

Dyrberg, N. R., Treusch, A. H., & Wiegand, C. (2016). Virtual laboratories in science education: students' motivation and experiences in two tertiary biology courses. *Journal of Biological Education*, 1–17. https://doi.org/10.1080/00219266.2016.1257498

Eliot, C. W. (1898). *Educational reform: essays and addresses*. New York: Century Co

Erdosne Toth, E., Morrow, B. L., & Ludvico, L. R. (2009). Designing blended inquiry learning in a laboratory context: A study of incorporating hands-on and virtual laboratories. *Innovative Higher Education*, 33(5), 333–344. https://doi.org/10.1007/s10755-008-9087-7

Fernández, G., and Morris, R. G. M. (2018). Memory, novelty and prior knowledge. *Trends in Neurosciences*, 41(10), 654–659. https://doi.org/10.1016/j.tins.2018.08.006

Furtak, E. M., Seidel, T., Iverson, H., and Briggs, D. C. (2012). Experimental and quasi-experimental studies of inquiry-based science teaching: A meta-Analysis. *Review of Educational Research*, 82(3), 300–329. https://doi.org/10.3102/0034654312457206

Geller, J., Carpenter, S. K., Lamm, M. H., Rahman, S., Armstrong, P. I., and Coffman, C. R. (2017). Prequestions do not enhance the benefits of retrieval in a STEM classroom. Cognitive Research: Principles and Implications, 2(1). https://doi.org/10.1186/s41235-017-0078-z

Glover, J. A. (1989). The "testing" phenomenon: not gone but nearly forgotten. *Journal of Educational Psychology*, 81(3), https://doi.org/10.1037/0022-0663.81.3.392

Godshalk, F. I., Swineford, F., & Coffman, W. E. (1966). *Measurement of Writing Ability* (C. E. E. Board (ed.); Research m). https://doi.org/10.1002/1520-6807(196701)4:1<93::AID-PITS2310040126>3.0.CO;2-Y

Gormally, C., Brickman, P., Brittan, H., and Armstrong, N. (2009). Effects of inquiry-based learning on students' science literacy skills and confidence. *International Journal for the Scholarship of Teaching and Learning*, 3(2). https://doi.org/10.20429/ijsotl.2009.030216

Goulden, N. R., and Griffin, C. J. G. (1995). The meaning of grades based on faculty and student metaphors. *Communication Education*, 44(2), 110–125. https://doi.org/10.1080/03634529509379003

Grainger, P., Adie, L., and Weir, K. (2016). Quality assurance of assessment and moderation discourses involving sessional staff. *Assessment and Evaluation in Higher Education*, 41(4), 548–559. https://doi.org/10.1080/02602938.2015.1030333

Griffin, D. K., Mitchell, D., & Thompson, S. J. (2009). Podcasting by synchronising PowerPoint and voice: What are the pedagogical benefits? *Computers and Education*, 53, 532–539. https://doi.org/10.1016/j.compedu.2009.03.011

Grömping, U. (2006). Relative importance for linear regression in R: The package relaimpo. *Journal of Statistical Software*, 17(1), 1–27. https://doi.org/10.18637/jss.v017.i01

Gross, D., Pietri, E. S., Anderson, G., Moyano-Camihort, K., Graham, M. J. (2015). Increased preclass preparation underlies student outcome improvement in the flipped classroom. *CBE Life Sciences Education*, https://doi.org/10.1187/cbe.15-02-0040

Gul, S., & Sozbilir, M. (2016). International trends in biology education research from 1997 to 2014: A content analysis of papers in selected journals. *Eurasia Journal of Mathematics, Science and Technology Education*, 12(6), 1631–1651. https://doi.org/10.12973/eurasia.2015.1363a

Gunstone, R. (1991). Reconstructing theory from practical experience. In B. E. Woolnough (Ed.), *Practical Science* (1st ed., pp. 67–77). Open University Press.

Gunstone, R. F., & Champagne, A. B. (1990). Promoting conceptual change in the laboratory. In E. Hegarty-Hazel (Ed.), *The student laboratory and the science curriculum.*

Guo, P. J., Kim, J., & Rubin, R. (2014). How Video Production Affects Student Engagement : An Empirical Study of MOOC Videos. *Proceedings of the 1st ACM Conference on Learning at Scale*, 41–50. https://doi.org/10.1145/2556325.2566239

Haagsman, M.E., Scager K., Boonstra, J., Koster, M.C. (2020). Pop-up questions within educational videos: effects on students' learning. *Journal of Science Education and Technology*, 29, 713-724. https://doi.org/10.1007/s10956-020-09847-3

Haagsman, M.E., Koster, M.C., Boonstra J., Scager K. (2021). Be prepared! How pre-lab modules affect students' understanding of gene mapping. *Journal of Science Education and Technology,* 30, 461-470. https://doi.org/10.1007/s10956-020-09890-0

Haagsman, M.E., Snoek B., Peeters A., Scager K. (2021). Examiners' use of rubric criteria for grading bachelor theses. *Assessment and Evaluation in Higher Education*. https://doi.org/10.1080/02602938.2020.1864287

Handelsman, J., Ebert-May, D., Beichner, R., Bruns, P., Chang, A., DeHaan, R., Gentile, J., Lauffer, S., Stewart, J., Tilghman, S. M., & Wood, W. B. (2004). Science teaching [6]. *Science*, 304(5670), 521–522. https://doi.org/10.1038/003208b0

Hannafin, M. J., and Hughes, C. W. (1986). A framework for incorporating orienting activities in computer-based interactive video. *Instructional Science*, 15(1), 239–255. https://doi.org/10.1007/BF00139613

Hapyak Interactive Video [Online application software]. (2020). Retrieved from https://corp.hapyak.com

Hawley, R., & Allen, C. (2019). Student-generated video creation for assessment: can it transform assessment within higher education? *International Journal for Transformative Research*, 5(1), 1–11. https://doi.org/10.2478/ijtr-2018-0001

Hihaho [Online application software]. (2020). Retrieved from https://hihaho.com

Hill, J. L., & Nelson, A. (2011). New technology, new pedagogy? Employing video podcasts in learning and teaching about exotic ecosystems. *Environmental Education Research*, 17(3), 393–408. https://doi.org/10.1080/13504622.2010.545873

Herreid, C. F., Schiller, N. A. (2013). Case studies and the flipped classroom. *Journal of College Science Teaching*, 42(5), 62–66.

Hofstein, A., and Lunetta, V. N. (2004). The laboratory in science education: foundations for the twenty-first century. Science Education, 88, 28–54. https://doi.org/10.1002/sce.10106

Hofstein, A., Navon, O., Kipnis, M., and Mamlok-Naaman, R. (2005). Developing students' ability to ask more and better questions resulting from inquiry-type chemistry laboratories. *Journal of Research in Science Teaching*, 42(7), 791–806. https://doi.org/10.1002/tea.20072

Hsin, W., & Cigas, J. (2013). Short Videos Improve Student Learning in Online Education. *Journal of Computing Sciences in Colleges*, 28(5), 253–259.

Huot, B. (1996). Toward a New Theory of Writing Assessment. *College Composition and Communication*, 47(4), 549–566.

Hutchings, P., & Shulman, L. S. (1999). The Scholarship of Teaching: New Elaborations, New Developments. *Change: The Magazine of Higher Learning*, 31(5), 10–15. https://doi.org/10.1080/00091389909604218

Huxley, T. (1899). *Science and education: Essays.* New York: Appleton and Company

H5P [Online application software]. (2020). Retrieved from https://h5p.org

Ibrahim, M. (2011). Implications of designing instructional video using cognitive theory of multimedia learning. *Critical Questions in Education*, 3(2), 83–104.

Jacoby, L. L., Wahlheim, C. N., Coane, J. H. (2010). Test-enhanced learning of natural concepts: Effects on recognition memory, classification, and metacognition. Journal of Experimental Psychology: *Learning Memory and Cognition*, 36(6), https://doi.org/10.1037/a0020636

Jing, H. G., Szpunar, K. K., and Schacter, D. L. (2016). Interpolated testing influences focused attention and improves integration of information during a video-recorded lecture. *Journal of Experimental Psychology: Applied*, 22(3), 305–318. https://doi.org/10.1037/xap0000087

Johnstone, A. H. (1997). Chemistry teaching — Science or alchemy? *Journal of Chemical Education*, 74(3), 262–268. https://doi.org/10.1021/ed074p262

Johnstone, A. H., and Al-Shuaili, A. (2001). Learning in the laboratory; some thoughts from the literature. C*hemistry Education*, 5(2), 42–51.

Johnstone, A. H., Sleet, R. J., and Vianna, J. F. (1994). An information processing model of learning: Its application to an undergraduate laboratory course in chemistry. *Studies in Higher Education*, 19(1), 77–87. https://doi.org/10.1080/03075079412331382163

Jones, S. M., and Edwards, A. (2010). Online pre-laboratory exercises enhance student preparedness for first year biology practical classes. *International Journal of Innovation in Science and Mathematics Education*, 18(2), 1–9.

Kapborg, I., and Berterö, C. (2002). Critiqing bachelor candidates' theses: Are the criteria useful? *International Nursing Review*, 49, 122–128. https://doi.org/10.1046/j.1466-7657.2002.00123.x

Kaplan, A. M., & Haenlein, M. (2016). Higher education and the digital revolution: About MOOCs, SPOCs, social media, and the cookie monster. *Business Horizons*, 59(4), 441–450. https://doi.org/10.1016/j.bushor.2016.03.008

Karpicke, J. D. (2009). Metacognitive Control and Strategy Selection: Deciding to Practice Retrieval During Learning. *Journal of Experimental Psychology: General*, 138(4), 469–486. https://doi.org/10.1037/a0017341

Karpicke, J. D., Blunt, J. R. (2011). Retrieval practice produces more learning than elaborative studying with concept mapping. *Science*, 331(February), 772–775.

Karpicke, J. D., and Roediger, H. L. (2007). Repeated retrieval during learning is the key to long-term retention. *Journal of Memory and Language*, 57(2), 151–162. https://doi.org/10.1016/j.jml.2006.09.004

Karpicke, J. D., Roediger, H. L. (2008). The critical Importance of retrieval for learning. *Science*, https://doi.org/10.1126/science.1152408

Kay, R. H. (2012). Exploring the use of video podcasts in education: A comprehensive review of the literature. *Computers in Human Behavior*, 28(3), 820–831. https://doi.org/10.1016/j.chb.2012.01.011

Kiley, M., & Mullins, G. (2004). Examining the examiners: How inexperienced examiners approach the assessment of research theses. *International Journal of Educational Research*, 41, 121–135. https://doi.org/10.1016/j.ijer.2005.04.009

Kirschner, P. A., Meester, M., Middelbeek, E., and Hermans, H. (1993). Agreement between student expectations, experiences and actual objectives of practicals in the natural sciences at the Open University of the Netherlands. *International Journal of Science Education*, 15(2), 175–197.

Kornell, N., Hays, M. J., and Bjork, R. A. (2009). Unsuccessful Retrieval Attempts Enhance Subsequent Learning. *Journal of Experimental Psychology: Learning Memory and Cognition*, 35(4), 989–998. https://doi.org/10.1037/a0015729

Kozma, R. (2003). The material features of multiple representations and their cognitive and social affordances for science understanding. *Learning and Instruction*, 13(2), 205–226. https://doi.org/10.1016/S0959-47520200021-X

Krajcik, J., Blumenfeld, P., Marx, R., and Soloway, E. (2000). Instruction, curricular, and technological supports for inquiry in science classrooms. In J. Minstrell and E. H. van Zee (Eds.), *Inquiring into Inquiry Learning and Teaching in Science* (pp. 283–315). American Association for the Advancement of Science.

Kiley, M., and Mullins, G. (2004). Examining the examiners: How inexperienced examiners approach the assessment of research theses. *International Journal of Educational Research*, 41, 121–135. https://doi.org/10.1016/j.ijer.2005.04.009

Kumar DD (2010) Approaches to interactive video anchors in problem-based science learning. *Journal of Science Education and Technology*, https://doi.org/10.1007/s10956-009-9154-6

Laaser, W., & Toloza, E. A. (2017). The changing role of the educational video in higher distance education. *International Review of Research in Open and Distance Learning*, 18(2), 264–276. https://doi.org/10.19173/irrodl.v18i2.3067

Landis, J.R., Koch, G. G. (1977). The Measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159. https://doi.org/10/2307/2529310

Lavigne, E., Risko, E. F. (2018). Optimizing the use of interpolated tests: The influence of interpolated test lag. Scholarship of *Teaching and Learning in Psychology*, https://doi.org/10.1037/stl0000118

Lawson, T. J., Bodle, J. H., Houlette, M. A., Haubner, R. R. (2006). Guiding questions enhance student learning from educational videos. *Teaching of Psychology*, https://doi.org/10.1207/s15328023top3301_7

Lax, N., Morris, J., Kolber, B. J. (2016). A partial flip classroom exercise in a large introductory general biology course increases performance at multiple levels. *Journal of Biological Education*, https://doi.org/10.1080/00219266.2016.1257503

Lee, J. L. C., Nader, K., and Schiller, D. (2017). An update on memory reconsolidation updating. *Trends in Cognitive Sciences*, 21(7), 531–545. https://doi.org/10.1016/j.tics.2017.04.006

Litman, J. A., Hutchins, T. L., and Russon, R. K. (2005). Epistemic curiosity, feeling-of-knowing, and exploratory behaviour. *Cognition and Emotion*, 19(4), 559–582. https://doi.org/10.1080/02699930441000427

Little, J. L., and Bjork, E. L. (2011). Pretesting with multiple-choice questions facilitate learning. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 33(33).

Lo, S. M., Gardner, G. E., Reid, J., Napoleon-Fanis, V., Carroll, P., Smith, E., & Sato, B. K. (2019). Prevailing questions and methodologies in biology education research: A longitudinal analysis of research in CBE-life sciences education and at the society for the advancement of biology education research. *CBE Life Sciences Education*, 18(1), 1–10. https://doi.org/10.1187/cbe.18-08-0164

Lo, C. K., Hew, K. F. (2017). A critical review of flipped classroom challenges in K-12 education: Possible solutions and recommendations for future research. *Research and Practice in Technology Enhanced Learning*, https://doi.org/10.1186/s41039-016-0044-2

Lowenthal, P. R., Borup, J., West, R. E., & Archambault, L. (2020). Thinking beyond Zoom: Using asynchronous video to maintain connection and engagement during the covid-19 pandemic. *Journal of Technology*, 28(2), 383–391. https://doi.org/10.24252/elite.v7i1a6

Lundgren, S. M., Halvarsson, M., and Robertsson, B. (2008). Quality assessment and comparison of grading between examiners and supervisors of bachelor theses in nursing education. *Nurse Education Today*, 28, 24–32. https://doi.org/10.1016/j.nedt.2007.02.009

Madlung, A., Bremer, M., Himelblau, E., & Tullis, A. (2011). A Study assessing the potential of negative effects in interdisciplinary math-biology instruction. *CBE Life Sciences Education*, 10(1), 43–54. https://doi.org/10.1187/cbe.10-08-0102

Makarevitch, I., and Kralich, E. (2011). Mapping maize genes: A series of research-based laboratory exercises. Biochemistry and Molecular Biology Education, 39(5), 375–383. https://doi.org/10.1002/bmb.20524

Makransky, G., Thisgaard, M. W., & Gadegaard, H. (2016). Virtual simulations as preparation for lab exercises: Assessing learning of key laboratory skills in microbiology and improvement of essential non-cognitive skills. *PLoS ONE*, 11(6), 1–11. https://doi.org/10.1371/journal.pone.0155895

Mayer, R. E. (1975). Forward transfer of different reading strategies evoked by test-like events in mathematics text. *Journal of Educational Psychology*, 67(2), 165–169. https://doi.org/10.1037/h0077011

Mayer, R. E. (2002). *Multimedia learning. The Psychology of Learning and Motivation*, 41, 85–139.

McComas, W. F. (2005). Laboratory instruction in the service of science teaching and learning. *The Science Teacher*, 72(7), 24–29.

McDaniel, M. A., Anderson, J. L., Derbish, M. H., and Morrisette, N. (2007). Testing the testing effect in the classroom. European *Journal of Cognitive Psychology*, 19(4–5), 494–513. https://doi.org/10.1080/09541440701326154

McIsaac, M. S., & Gunawardena, C. N. (2001). Distance education. In *The handbook of research for educational communities and technology* (pp. 410–415). The Association for Educational Communications and Technology. https://doi.org/10.1016/B978-0-12-813467-2.00063-8

Minner, D. D., Levy, A. J., and Century, J. (2010). Inquiry-based science instruction-what is it and does it matter? Results from a research synthesis years 1984 to 2002. *Journal of Research in Science Teaching*, 47(4), 474–496. https://doi.org/10.1002/tea.20347

Mutch-Jones, K., Sengupta, N., Minor, V. C., & Goudsouzian, L. K. (2021). Professional science education videos improve student performance in nonmajor and intermediate biology laboratory courses. *Biochemistry and Molecular Biology Education*, 49, 151–159. https://doi.org/10.1002/bmb.21415

Nadelson, L. S., Scaggs, J., Sheffield, C., and Mcdougal, O. M. (2015). Integration of video-based demonstrations to prepare students for the organic chemistry laboratory. Journal of Science Education and Technology, 24, 476–483. https://doi.org/10.1007/s10956-014-9535-3

National Research Council. (2012). Discipline-based education research: Understanding and improving learning in undergraduate science and engineering (S. R. Singer, N. R. Nielsen, & H. A. Schweingruber (eds.)). The National Academies Press. https://doi.org/10.17226/13362

National Society for the Study of Education. (1932). A *Program for Teaching Science: Thirty-first Yearbook of the NSSE.*

Offerdahl, E. G., Balser, T., Dirks, C., Miller, K., Momsen, J. L., Montplaisir, L., Osgood, M., Sirum, K., Pat Wenderoth, M., White, B., Wood, W. B., Withers, M., & Wright, R. (2011). Society for the advancement of biology education research (SABER). *CBE Life Sciences Education*, 10(1), 11–13. https://doi.org/10.1187/cbe.10-11-0135

Panadero, E., Alonso-Tapia, J., and Reche,E. (2013). Rubrics vs. self-assessment scripts effect on self-regulation, performance and self-efficacy in pre-service teachers. *Studies in Educational Evaluation*, 39, 125–132. https://doi.org/10.1016/j.stueduc.2013.04.001

PanoptoTM [Online application software]. (2020). Retrieved from https://www.panopto.com

Pastötter, B., Bäuml, K. H. T. (2014). Retrieval practice enhances new learning: The forward effect of testing. *Frontiers in Psychology*, https://doi.org/10.3389/fpsyg.2014.00286

Pedersen, T. L. (2019). *patchwork: The composer of plots* (R-package version 1.0.0). https://cran.r-project.org/package=-patchwork

Playposit [Online application software]. (2020). Retrieved from https://go.playposit.com

Prins, F. J., de Kleijn, R., and van Tartwijk, J. (2017). Students' use of a rubric for research theses. *Assessment & Evaluation in Higher Education*, 42(1), 128–150. https://doi.org/10.1080/02602938.2015.1085954

Pogacnik, L., and Cigic, B. (2006). How to motivate students to study before they enter the lab. *Journal of Chemical Education*, 83(7), 1094–1098. https://doi.org/10.1021/ed083p1094

Pols, F. (2020). A physics Lab Course in times of COVID-19. *Electronic Journal for Research in Science & Mathematics Education*, 24(2), 172–178.

Pyc, M. A., and Rawson, K. A. (2009). Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory and Language*, 60(4), 437–447. https://doi.org/10.1016/j.jml.2009.01.004

Ray, S., & Srivastava, S. (2020). Virtualization of science education: a lesson from the COVID-19 pandemic. *Journal of Proteins and Proteomics*, 11(2), 77–80. https://doi.org/10.1007/s42485-020-00038-7

Reddy, Y. M., and Andrade, H. (2010). A review of rubric use in higher education. *Assessment & Evaluation in Higher Education*, 35(4), 435–448. https://doi.org/10.1080/02602930902862859

&

Reid, N., and Shah, I. (2007). The role of laboratory work in university chemistry. *Chemistry Education Research and Practice*, 8(2), 172–185. https://doi.org/10.1039/B5RP90026C

Revelle, W. (2019). *psych: Procedures for psychological, psychometric, and personality Research* (R package version 1.9.12). https://cran.r-project.org/package=-psych

Richland, L. E., Kornell, N., and Kao, L. S. (2009). The pretesting effect: do unsuccessful retrieval attempts enhance learning? *Journal of Experimental Psychology: Applied*, 15(3), 243–257. https://doi.org/10.1037/a0016496

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

Roediger, H. L., Karpicke, J. D. (2006a). The power of testing memory; Basic research and implications for educational practice. *Perspectives on Psychological Science*, 1(3), 181–210.

Roediger, H. L., and Karpicke, J. D. (2006b). Test-Enhanced Learning Taking Memory Tests Improves Long-Term Retention. *Psychological Science*, 17(3), 249–255. https://doi.org/10.1111/j.1467-9280.2006.01693.x

Rothkopf, E. Z., and Coke, E. U. (1963). Repetition interval and rehearsal method in learning equivalences from written sentences. *Journal of Verbal Learning and Verbal Behavior*, 2(5–6), 406–416. https://doi.org/10.1016/S0022-5371(63)80041-9

Rouet, J., Vidal-Abarca, E., Erboul, A. B. (2001). Effects of information search tasks on the comprehension of instructional text. *Discourse Processes*, https://doi.org/10.1207/S15326950DP3102

Sadler, D. R. (2009). Indeterminacy in the use of preset criteria for assessment and grading. *Assessment and Evaluation in Higher Education*, 34(2), 159–179. https://doi.org/10.1080/02602930801956059

ScalableLearning [Online application software]. (2020). Retrieved from https://www.scalable-learning.com/#/home

Schacter, D. L., and Szpunar, K. K. (2015). Enhancing attention and memory during video-recorded lectures. *Scholarship of Teaching and Learning in Psychology*, 1(1), 60–71. https://doi.org/10.1037/st10000011

Schmid, S., and Yeung, A. (2005). The influence of a pre-laboratory work module on student performance in the first year chemistry laboratory. *HERDSA Annual Conference*, 471–479.

Schofer, E., & Meyer, J. W. (2005). The worldwide expansion of higher education in the twentieth century. *American Sociological Review*, 70(12), 898–920. https://doi.org/10.1177/000312240507000602

Schroeder, C. M., Scott, T. P., Tolson, H., Huang, T.-Y., and Lee, Y.-H. (2007). A meta-analysis of national research: Effects of teaching strategies on student achievement in science in the United States. *Journal of Research in Science Teaching*, 44(10), 1436–1460. https://doi.org/10.1002/tea

Schweppe, J., and Rummer, R. (2014). Attention, working memory, and long-term memory in multimedia learning: an integrated perspective based on process models of working memory. *Educational Psychology Review*, 26(2), 285–306. https://doi.org/10.1007/s10648-013-9242-2

Singer, S. R., Nielsen, N. R., & Schweingruber, H. A. (2011). D*iscipline-based education research: Understanding and Improving Learning in Undergraduate Science and Engineering*. The National Academies Press.

Slemmons, K., Anyanwu, K., Hames, J., Grabski, D., Mlsna, J., Simkins, E., & Cook, P. (2018). The Impact of Video Length on Learning in a Middle-Level Flipped Science Setting: Implications for Diversity Inclusion. *Journal of Science Education and Technology*, 27(5), 469–479. https://doi.org/10.1007/s10956-018-9736-2

Smith, A., & Hall, E. (1902). *The teaching of chemistry and physics in the secondary school.*

Stockwell, B. R., Stockwell, M. S., Cennamo, M., & Jiang, E. (2015). Blended Learning Improves Science Education. *Cell*, 162(August), 933–936. https://doi.org/10.1016/j.cell.2015.08.009

Sweller, J., Van Merrienboer, J. J. G., and Paas, F. G. W. C. (1998). Cognitive architecture and instructional design. *Educational Psychology Review*, 10(3), 251–296. https://doi.org/10.1023/A:1022193728205

Szpunar, K. K., Jing, H. G., Schacter, D. L. (2014). Overcoming overconfidence in learning from video-recorded lectures: Implications of interpolated testing for online education. *Journal of Applied Research in Memory and Cognition*, https://doi.org/10.1016/j.jarmac.2014.02.001

Szpunar, K. K., Khan, N. Y., Schacter, D. L. (2013). Interpolated memory tests reduce mind wandering and improve learning of online lectures. *Proceedings of the National Academy of Sciences*, https://doi.org/10.1073/pnas.1221764110

Szpunar, K. K., McDermott, K. B., Roediger, H. L. (2008). Testing during study insulates against the buildup of proactive interference. *Journal of Experimental Psychology: Learning Memory and Cognition*, https://doi.org/10.1037/a0013082

Tai, J., Ajjawi, R., Boud, D., Dawson, P., and Panadero, E. (2018). Developing evaluative judgement: enabling students to make decisions about the quality of work. *Higher Education*, 76, 467–481. https://doi.org/10.1007/s10734-017-0220-3

Timmerman, B. E., Strickland, D. C., Johnson, R. L., and Payne, J. R. (2011). Development of a ‘universal’ rubric for assessing undergraduates’ scientific reasoning skills using scientific writing. *Assessment & Evaluation in Higher Education*, 36(5), 509–547. https://doi.org/10.1080/02602930903540991

Thomas, D.R. (2006). A general inductive approach for analyzing qualitative evaluation data. *American Journal of Evaluation*, 27(2), 237-246. http://doi.org/10.1177/1098214005283748

Tight, M. (2018). Tracking the Scholarship of Teaching and Learning. *Policy Reviews in Higher Education*, 2(1), 61–78. https://doi.org/https://doi.org/10.1080/23322969.2017.1390690

Timmerman, B. E., Strickland, D. C., Johnson, R. L., & Payne, J. R. (2011). Development of a ‘universal’ rubric for assessing undergraduates’ scientific reasoning skills using scientific writing. *Assessment & Evaluation in Higher Education*, 36(5), 509–547. https://doi.org/10.1080/02602930903540991

Tobi, H., & Kampen, J. K. (2018). Research design: the methodology for interdisciplinary research framework. *Quality and Quantity*, 52(3), 1209–1225. https://doi.org/10.1007/s11135-017-0513-8

Toftness, A. R., Carpenter, S. K., Lauber, S., and Mickes, L. (2018). The Limited effects of prequestions on learning from authentic lecture videos. *Journal of Applied Research in Memory and Cognition*, 7(3), 370–378. https://doi.org/10.1016/j.jarmac.2018.06.003

Tsai, C. C., & Wen, M. L. (2005). Research and trends in science education from 1998 to 2002: A content analysis of publication in selected journals. International *Journal of Science Education*, 27(1), 3–14. https://doi.org/10.1080/0950069042000243727

Tulving, E. (1967). The effects of presentation and recall of material in free-recall learning. *Journal of Verbal Learning and Verbal Behavior*, 6(2), 175–184. https://doi.org/10.1016/S0022-5371(67)80092-6

Van der Kolk, K., Beldman, G., Hartog, R., and Gruppen, H. (2012). Students using a novel web-based laboratory class support System: A case study in food chemistry education. *Journal of Chemical Education*, 89, 103–108. https://doi.org/10.1021/ed1005294

van der Vleuten, C. P. M., Schuwirth, L. W. T., Driessen, E. W., Dijkstra, J., Tigelaar, D., Baartman, L. K. J., and van Tartwijk, J. (2012). A model for programmatic assessment fit for purpose. *Medical Teacher*, 34(3), 205–214. https://doi.org/10.3109/0142159X.2012.652239

Van Kesteren, M. T. R., Rijpkema, M., Ruiter, D. J., and Fernández, G. (2010). Retrieval of associative information congruent with prior knowledge is related to increased medial prefrontal activity and connectivity. *Journal of Neuroscience*, 30(47), 15888–15894. https://doi.org/10.1523/JNEUROSCI.2674-10.2010

Vural, O. F. (2013). The impact of a question-embedded video-based Learning tool on e-learning. *Educational Sciences: Theory Practice*, 13(2), 1315–1323.

Warnes, G. R., Bolker, B., Bonebakker, L., Gentleman, R., Huber, W., Liaw, A., Lumley, T., Maechler, M., Magnusson, A., Moeller, S., Schwartz, M., Venables, B., and Galili, T. (2019). *gplots: Various R programming tools for plotting data* (3.0.3). https://github.com/talgalili/gplots

Wei, C. A., and Woodin, T. (2011). Undergraduate research experiences in biology: Alternatives to the apprenticeship model. *CBE Life Sciences Education*, 10(2), 123–131. https://doi.org/10.1187/cbe.11-03-0028

Whittle, S. R., and Bickerdike, S. R. (2015). Online preparation resources help first year students to benefit from practical classes. *Journal of Biological Education*, 49(2), 139–149. https://doi.org/10.1080/00219266.2014.914554

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag. https://doi.org/978-3-319-24277-4

Wieling, M. B., Hofman, W. H. A. (2010). The impact of online video lecture recordings and automated feedback on student performance. *Computers and Education*, https://doi.org/10.1016/j.compedu.2009.10.002

Wiggins, G. P. (1993). *Assessing student performance: exploring the purpose and limits of testing* (Jossey-Bass (ed.)).

Wijnker, W. (2021). The Unseen Potential of Film for Learning. Film's Interest Raising Mechanisms Explained in Secondary Science and Mathematics Education. Utrecht: Utrecht University.

Wilke, C. O. (2019). *cowplot: Streamlined plot theme and plot annotations for ggplot2*. https://github.com/wilkelab/cowplot

Williams, L., and Kemp, S. (2019). Independent markers of master's theses show low levels of agreement. *Assessment and Evaluation in Higher Education*, 44(5), 764–771. https://doi.org/10.1080/02602938.2018.1535052

Winberg, M. T., and Berg, A. R. (2007). Students' cognitive focus during a chemistry laboratory exercise: effects of a computer-stimulated prelab. *Journal of Research in Science Teaching*, 44(8), 1108–1133. https://doi.org/10.1002/tea

Woodfield, B. F., Catlin, H. R., Waddoups, G. L., Moore, M. S., Swan, R., Allen, R., & Bodily, G. (2004). The Virtual ChemLab project: A realistic and sophisticated simulation of inorganic qualitative analysis. *Journal of Chemical Education*, 81(11), 1672–1678. https://doi.org/10.1021/ed082p1728

Zhang, L. (2016). Is inquiry-based science teaching worth the effort? Some thoughts worth considering, 25, 897–915. https://doi.org/10.1007/s11191-016-9856-0

&

Nederlandse samenvatting

Onderwijs is continu aan verandering onderhevig. Zo verandert de leeromgeving of het aantal studenten, worden andere vaardigheden van belang, komen er technische onderwijstools beschikbaar of weten we simpelweg beter hoe of wanneer studenten het beste leren. Hierdoor is er een continue vraag om te bestuderen naar hoe studenten leren, wanneer studenten het meest leren en hoe we ons onderwijs het beste vorm kunnen geven. In dit proefschrift worden een aantal ontwikkelingen binnen het onderwijs besproken en wordt bestudeerd welk effect deze ontwikkelingen hebben op het leerproces van studenten. De vragen van deze studies komen direct uit het veld, in dit geval uit het onderwijs van het Departement Biologie aan de Universiteit Utrecht.

Hoofdstuk 2

In de eerste twee hoofdstukken wordt het gebruik van interactieve video's besproken. Deze studies zijn uitgevoerd in de eerstejaars cursus moleculaire biologie. Tijdens deze cursus kijken studenten wekelijks naar video's waarin onderwerpen over de moleculaire biologie worden uitgelegd. Daarna maken ze thuis een kennistoets om te controleren of ze de stof hebben begrepen, maken ze verdiepende opdrachten op de campus en bespreken ze de leerstof en opdrachten na met de docent.

De video's van deze cursus bevatten een aantal quizvragen om ervoor te zorgen dat de studenten meteen actief bezig zijn met het verwerken van de stof. Deze vragen worden door ons ook wel "pop-up vragen" genoemd omdat ze tijdens het bekijken van de video plots in beeld verschijnen en dan direct moeten worden beantwoord. In de eerste studie van hoofdstuk 2 worden twee groepen studenten met elkaar vergeleken: een groep die video's bekeek met pop-up vragen en een groep die dezelfde video's bekeek zonder pop-up vragen. Beide groepen deden vervolgens een test over de besproken leerstof in de video's. Hierbij bleek dat de studenten gemiddeld een hogere toets score hadden indien ze pop-up vragen moesten beantwoorden. In de tweede studie bekeken twee groepen wederom dezelfde video's maar beantwoordde de ene groep pop-up vragen over het ene onderwerp en de andere groep over een ander onderwerp. Als de studenten vervolgens weer een test deden, bleek dat de ene groep niet beter scoorde op een bepaald onderwerp dan de andere groep. Kortom, deze twee studies suggereren dat niet persé de inhoud, maar enkel al de aanwezigheid van een pop-up vraag effect heeft op de leeropbrengst van een video. We adviseren docenten om pop-up vragen in hun video's te brengen opdat studenten, alleen al door de aanwezigheid van de vraag, meer leerstof uit de video zullen onthouden en begrijpen.

Om uit te zoeken hoe de pop-up vragen bijdragen aan de leeropbrengst van de student, vroegen we de studenten wat ze doen als ze zo'n pop-up vraag beantwoorden. Zo'n 37% van de studenten gaf aan dat ze, indien ze het antwoord op de vraag niet wisten, de video terugspoelden om het antwoord op te zoeken. Dit terugspoel gedrag wordt ook bevestigd uit kijkersdata. Eén van de effecten van pop-up vragen is dus dat studenten actief bepaalde delen van een video gaan terugkijken.

Hoofdstuk 3

In de studies uit hoofdstuk 2 zijn pop-up vragen gebruikt die testen of de studenten de voorgaande stof uit de video hebben begrepen. Je kan echter ook pop-up vragen gebruiken om studenten alvast te laten nadenken over de leerstof die komen gaat. Beide type vragen, post-vragen en pre-vragen, werden in de eerstejaars cursus moleculaire biologie gebruikt. De verwachting was dat andere leerprocessen plaatsvinden als er een pre-vraag wordt gesteld dan als er een post-vraag wordt gesteld. In hoofdstuk 3 wordt besproken welke leerprocessen er volgens de studenten plaatsvinden als zij pre-vragen of post-vragen beantwoorden. Om deze leerprocessen te onderzoeken, werden de studenten in drie groepen verdeeld. De eerste groep bekeek een video met pre-vragen en de tweede groep bekeek dezelfde video met post-vragen. Een derde groep bekeek de video met vragen die, net als bij de eerste groep, vooraf aan het bijbehorende videofragment verschenen maar die ze enkel na dit videofragment hoefden te beantwoorden. Vervolgens werden de groepen gevraagd wat zij van de pop-up vragen vonden. Alle drie groepen gaven veelvuldig aan dat ze door de pop-up vragen beter hun aandacht konden vasthouden bij het bekijken van de video's. Ook gaven studenten in alle groepen regelmatig aan dat ze zo hun begrip konden testen. De studenten die enkel pre-vragen beantwoorden, zeiden echter ook dat ze door de pop-up vragen hun voorkennis konden testen en ophalen. Daarnaast benoemden de studenten uit de derde groep nog een ander bijkomend voordeel: zij zeiden dat het stellen van pop-up vragen aangaf wat belangrijk is om te weten en dat ze daardoor wisten waar ze op moesten letten tijdens het bekijken van een video. Kortom, de antwoorden van de studenten suggereren dat het stellen van pre-vragen, post-vragen of een combinatie van beiden, inderdaad verschillende effecten hebben op het leerproces van de student.

Om te onderzoeken of de verschillende type vragen ook andere effecten hebben op het leerbegrip en het onthouden van de stof, werden de studenten ook na het bekijken van de video getest op de informatie uit de video. Hieruit bleek dat de testresultaten niet of nauwelijks verschilden tussen deze groepen. Wij geven daarom geen algemene voorkeur voor het stellen van pre-vragen en/of post-vragen maar adviseren docenten om pre-vragen voornamelijk in te zetten als je voorkennis wilt testen en post-vragen in te zetten als je wilt dat studenten de stof uit de video verwerken. Het vooraf meegeven van vragen die achteraf moeten worden beantwoord wordt aangeraden indien je wilt dat studenten hun aandacht op een specifiek onderwerp vestigen.

Hoofdstuk 4

Hoofdstuk 4 gaat over practicumonderwijs en is uitgevoerd in de tweedejaars cursus moleculaire genetische onderzoekstechnieken. Tijdens deze cursus doen de studenten experimenten en verwerken ze de aanleiding, methode en hun bevindingen in een lab rapport. Veel experimenten in de biologie hebben als hoofddoel dat studenten de achtergrondtheorie beter begrijpen en deze informatie ook kunnen koppelen aan de praktijk. Uit eerdere studies blijkt echter dat studenten zich tijdens een experiment nauwelijks bezighouden met de achterliggende theorie en zich voornamelijk richten op het stapsgewijs volgen van het protocol. Een mogelijke verklaring hiervoor is dat studenten tijdens een experiment op zoveel dingen moeten letten waardoor ze ervoor kiezen om zich louter te richten op het opvolgen van de experimentele stappen. Om ervoor te zorgen dat studenten toch die koppeling tussen de theorie en praktijk maken, is er een computermodule ontwikkelt die studenten vooraf aan het experiment doorlopen. Zo'n activiteit ter voorbereiding van een experiment wordt ook wel een pre-lab activiteit genoemd. De pre-lab uit deze studie bestaat uit een inleiding met animaties over de achtergrondtheorie van het experiment, gevolgd door theoretische vragen over het protocol en het interpreteren van data. In de studie uit hoofdstuk 4 werden twee groepen uit opeenvolgende jaren met elkaar vergeleken: een groep die geen pre-lab deed en een groep die wel een pre-lab deed. Vervolgens is zowel voor, tijdens als na het experiment gemeten in hoeverre studenten de achtergrondtheorie begrepen en deze konden koppelen aan de praktijk. Als eerste werden studenten bij binnenkomst op de labzaal gevraagd om een toets te doen over de achtergrondtheorie. Zoals verwacht werd deze toets significant beter gemaakt door de studenten die vooraf de pre-lab hadden gedaan. Ten tweede werd er gekeken waar studenten zich tijdens het experiment mee bezighielden door alle vragen op te nemen die zij aan hun docent stelden. Alhoewel studenten in beide groepen nauwelijks vragen stelden over de achtergrondtheorie was er wel één duidelijk verschil: studenten die de pre-lab hadden gedaan stelden minder vaak vragen over de procedure van het experiment, wat suggereert dat deze studenten beter wisten wat er van hen verwacht werd. Tot slot werd er ook gekeken of deze studenten na afloop van het experiment beter waren in het koppelen van de theorie aan de praktijk. Hiervoor werden hun labrapporten beoordeeld op basis van begrip. De studenten die de pre-lab module hadden gedaan bleken hierbij significant hogere scores te halen voor zowel de inleiding, resultaten, discussie als conclusie. Dit suggereert dat de pre-lab module inderdaad bijdraagt aan het begrip van de achtergrondtheorie en hoe dit relateert aan de praktijk.

Hoofdstuk 5

Hoofdstuk 5 gaat over het beoordelen van bachelorverslagen. In Nederland moeten studenten vaak verplicht een bachelorverslag schrijven aan het einde van hun studie, zo ook tijdens de bachelor biologie aan de Universiteit Utrecht. Voor dit bachelorverslag moeten studenten een review schrijven van bestaande primaire literatuur. Studenten worden daarbij onder andere geacht om zelf relevante primaire literatuur te zoeken, de informatie uit de literatuur kritisch te bekijken, mogelijke limitaties en alternatieve verklaringen te benoemen, goed onderbouwde conclusies te maken, de relevantie van het onderzoek aan te stippen en hier een duidelijk en gestructureerd verhaal van te schrijven. Kortom, studenten worden geacht om aan een lange lijst van criteria te voldoen. Dit betekent ook dat al deze criteria in acht moeten worden genomen bij het beoordelen van een bachelorverslag, terwijl er uiteindelijk maar één globaal cijfer aan het verslag wordt toegekend. Om beoordelaars te begeleiden in dit beoordelingsproces worden zij geacht om een rubric in te vullen. Dit is een document waarin de verschillende beoordelingscriteria van het bachelor verslag staan beschreven, en waarin voor elk criterium staat beschreven wanneer deze als onvoldoende, voldoende of goed wordt beschouwd. Een beoordelaar kan zo aan een student voor elk criterium feedback geven. Er bestaat echter geen formule waarmee het eindcijfer op basis van deze criteria scores kan worden bepaald. De beoordelaar staat dus vrij om te bepalen hoe hij of zij de rubric gebruikt om tot het uiteindelijke eindcijfer te komen. In de studie in hoofdstuk 5 is de relatie tussen de criteria scores en het uiteindelijke eindcijfer van 318 bachelor verslagen onderzocht. Hieruit blijkt dat er per eindcijfer een grote variatie is in de gemiddelde rubric score die door de beoordelaar is toegekend. Een mogelijke verklaring hiervoor is dat sommige criteria meer worden meegewogen dan andere. Inderdaad blijkt bij verdere analyse dat rubric scores voor wetenschappelijke kwaliteit en structuur het meest correleren met eindcijfer, terwijl rubric scores voor titel, spelling en grammatica het minst correleren met het eindcijfer. Dit komt ook overeen met de criteria die door het departement als meest of minst relevant worden beschouwd. De analyses van deze studie geven inzicht in de criteria die beoordelaars het meest relevant vinden voor het schrijven van een bachelor verslag. Departementen worden aangemoedigd om, ter kwaliteitscontrole, zulke analyses van tijd tot tijd uit te voeren om te evalueren of de eindcijfers een goede afspiegeling zijn van het uiteindelijke product en of verslagen voornamelijk worden beoordeeld op de criteria die ook door het departement als meest relevant worden beschouwd.

&

&

About the author

Marjolein Haagsman was born on August 4, 1987, in San Francisco, CA, USA. In 1999 she started her pre-university education at De Breul in Zeist, from which she graduated at Atheneum level in 2006, on the subjects of Dutch, English, Mathematics B, Biology, Physics, Chemistry and Visual Arts. In 2006, she started as a student in the bachelor Biology at Utrecht University. She received her bachelor's degree *cum laude* in 2009. Marjolein continued her education with a master's in Molecular and Cellular Life Sciences at Utrecht University. During her master she investigated the breakdown and formation of the Golgi apparatus during cell division. She also performed research at the Patel lab at UC Berkeley where she studied the regeneration of germline cells in *Parhyale Hawaiensis*. Marjolein received her master's degree in 2011 with an 8.4 on average (GPA of 4.0).

During her studies, Marjolein developed interests in education while being a tutor for pre-university students and a teacher assistant for undergraduates. In 2011, she started a masters' in education at Utrecht University. Marjolein soon decided to switch to the master U-TEAch that is specified in international and bilingual education. During these masters, Marjolein did internships at the St Bonifatius College in Utrecht, Slingerbos/Levant College in Harderwijk, Slingerbos/Levant College in Zeewolde and the Pretoria High School for Girls in Pretoria, South Africa. Marjolein graduated her master's in international and bilingual education in 2013. In 2013, Marjolein became a science teacher for School at Sea where she taught chemistry, biology, mathematics and science to students aged 15-17 while sailing across the Caribbean and Atlantic Ocean. In 2014, Marjolein started to work at Utrecht University to develop course material and evaluate courses of the new bachelor track Molecular Life Sciences. At the same time, she was teaching biology and science for students aged 11-14 at the International School in Almere. In 2016, Marjolein started a PhD in Biology education at the Department of Biology at Utrecht

University. During her PhD, she has been a teacher in Molecular Biology and a teacher and course coordinator in Biotechnology. Marjolein also obtained her Basic Qualification in Education (Basiskwalificatie Onderwijs; BKO), became an organizing member of Special Interest Groups in both video education and learning spaces, and worked as a video-expert as part of the educational program Educate-it of Utrecht University. She currently works as a lecturer and educational developer at the Department of Biology at Utrecht University.

Acknowledgments

This journey started many years ago, when I was only sixteen years old. My mathematics teacher (*meneer Veerkamp, Lex*) asked me if I could tutor one of his students in mathematics. That's when I became an enthusiast in teaching, and I never stopped ever since. This once again shows how important teachers can be. Meneer Veerkamp: Thanks alot. Hopefully I can inspire my students like you once did for me!

Secondly, I would like to thank one of the most respected teachers we ever had at the biology department: *Fons Cremers*. Fons, you were my teacher when I was still a student, you were my boss when I was your student assistant and my colleague when we were teaching together. I got to know you as an enthusiastic and caring teacher and learned that these are one of the most important characteristics of a good teacher. Thanks for teaching me how to teach!

And of course, I would like to give many special thanks to my promotor and co-promotors. *Johannes*: I am very grateful that you wanted to become my supervisor, even though you already became professor emeritus after my first year. I love your Frisian no-nonsense way of thinking and felt supported by you as we both had some trouble in getting used to the differences between biological and educational research. I would still love to see your lambs in spring! *Margot*: of all my colleagues I had most contact with you. You were not only my supervisor, but also my fellow teacher in Biotechnology. We both share a great feeling of responsibility and perfection when it comes to education, and that is probably why we are a great team. Luckily, we will still keep on working together! *Karin*: I did not know you when I started my PhD, but Johannes claimed that you would be a great asset to the team, and he was definitely right. Your expertise in education and educational research was much needed and I highly appreciate your patience and great help on short notices. Johannes, Margot and Karin: thank you all for all our sessions, reading many proofs, coding lots of student responses and grading numerous lab reports. I am grateful for all of your help and efforts!

I would also like to thank *Han Wösten* for giving me the trust and opportunity to perform this unique PhD project as part of our education institute. Also, thanks to *Anna Akhmanova* and *Casper Hoogenraad*, who gave me the opportunity to develop and revise Molecular Life Sciences, which is how I enrolled in this PhD project. You might still remember how surprised I was when you proposed to do a PhD in education, but I am glad I said yes!

With that I would also thank all the other great members of the Cell Biology labs for adopting me and for all the fun lunches, lectures, borrels, parties and outings we had. Thank you so much, *Lukas, Paul, Florian, Ginny, Esther, Jan Andries, Harold, Corette, Peter-Jan, René, Ilya, Bart, Eugene, Phebe, Anna, Mithila, Amélie, Yolanda, Martin, Arthur, Joyce, Funso, Josiah, Marvin, Boris, Kyle, Babet, Cyntha, Robin, Lisa, Fangrui, Hugo, Lotte, Jessica, Malina, Klara, Daphne, Robbelien, Giel, Sybren, Dipti, Nicky, Marijn, Manon, Jelmer, Catherina, Gabriela, Olga, Inês, Maud, Desiree, Ruddi, Anna, Josta, Petra, Marleen, Bas, Sam, Marta, Catia, Ivar, Amol, Dieudonnee, Jian, Yujie, Anne, Hai Yin, Roderick, Feline, Xingxiu* and *Sofia*. My PhD project wouldn't be as fun without you. A special thanks to all of my former O522 office roommates: *Frits, Astrid, Ben, Dennis, Wilco, Chao, Elske, Laura, Eitan, Juliane, Suzanne, João, Sabrina, Marvin* and *Carlijn*. I had a great time with you in our huge office space, although it's too bad that the office piano and couch never became reality. Thank you *Astrid* for our nice hikes in Amelisweerd during our breaks. And of course, a special thanks to my grumpy office mate *Dennis*: You contributed to the content of this thesis as you helped me with my first figures and illustrator skills, but most of all: you provided me with emergency chocolate and distracted me with weird news articles when needed. I had great fun with you, and I definitely missed you during my last year. But now you are also working in education so in fact we are still colleagues! Unfortunately, I could only spend little time in my new office in N505. Roomies *Lena, Ate and Ma*x: thank you for your warm welcome to the office!

A special thanks to *Laurens* and *Ron*. It is great to teach the very first course of the bachelors with you and I appreciate that you were willing to participate in studies on our molecular biology course. We are a great team! All other *teachers at the biology department* with whom I discuss education in the

hallway or during our monthly meetings: thank you for inspiring me with your experiences! All dedicated and supportive staff of *OWI*: thanks for being such a great support to our education. A special thanks to *Elly* who helped develop the pre-lab module that is part of the study in chapter 4.

I would also like to thank *Martijn* for providing the opportunity to work on his large dataset of thesis assessments. Thanks also to *Basten*, *Ton* and *Frans* who helped to turn this study in a grounded and nicely visualized paper. *Marcel* and other *MGOT teachers*: thank you for participating in our study on pre-labs and willingness to record your lessons. I would also like to thank *Bas* and the rest of the *USO lab project members* for sparking my interest in lab education. I still believe that there is a lot to improve when it comes to laboratory education and the coming time, I will keep myself busy on improving lab education at the Department of Biology.

Being the only PhD student in education at the department, I was lucky to be welcomed by other educational research groups. First of all, I would like to thank the *expertise centre of education at UMC Utrecht* for their warm welcome. Special thanks to *Rianne*, who immediately invited me to their work sessions and meetings and made me feel welcome to join. Second, I would like to thank *Toine Pieters* and the others of the *Freudenthal Institute* for giving me the opportunity to join their Thursday meetings to get inspired and share my thoughts with the FI team. A special thanks to *Winnifred*: it was nice to share our frustrations and thoughts on educational videos and I learned a lot from your perspectives on film.

This PhD project only started once we were inspired by *Jasper* to use interactive videos, got support from Educate-it to implement these in our own education and got help from *Renske* to evaluate how our students were viewing our videos. Thank you, *Jasper*, for inspiring us and thousands of other educators with your awesome and creative ideas on education. Thank you, *Jan*, *Mabelle* and the others of the *Educate-it team* for being of such a great help. I've seen how Utrecht University is at the forefront when it comes to innovations on education, and you've definitely been an important factor in this. I very much enjoy being part of your creative and inspiring team and thank you all for the great borrels, game evenings and Christmas dinners!

Then certainly not to forget: our many great *students*. You make my days of teaching a great fun! Besides, without you this thesis wouldn't exist. Thank you all for sharing your interesting ideas, suggestions and experiences on our education. You keep inspiring me!

And finally, I would like to give my final thanks to all my friends and family for the great times in between. First of all: Gill, Iek, Lin, Tess and Bet (*Gillian, Ike, Linde, Tessa and Annebet*); I love our chaotic, random and loving discussions at our amazing diners. A special thanks to my best friend *Annebet*: you are the greatest person to spend my weekends with! And to all of the other awesome biologists: *Esmée, Marjolijn, Michiel, Anique, Anneke and many other biologists, oud-UBV'ers* and *alumni-veldweekenders*: you are simply the best! *Computer scientists:* thanks for playing many board games. *U-TEACHers*: you have been at the basis of my career in education, and it was fantastic. *Gerrit-Jan*, *Frans* and *Sjaak* already predicted I would go back into science someday. I said they were wrong, but here I am!

Annika and *Karlijn*: my paranymphs, grote zussen and best friends! I am so lucky to have two great sisters as paranymphs. We celebrated both papa's inauguration and mama's defense at the Academiegebouw. It is great to also celebrate this defense with you, once again at Domplein. I look forward to all of our nice travels, zussenweekenden and many other fun activities. And Karlijn: as a biologist, I was lucky to have a sister with a background in social sciences. Thanks for introducing me to your research techniques and statistical analyses.

Pap and *Mam*; You have taught me from early on what science is about. Our discussions at the dining table were, and still are, on literature, culture, science, biology and chemistry. Thank you for teaching us to be open-minded and think critically, but most of all: thank you for being such a great support, always.

Maarten; you were and are the greatest support of all, especially in the last year when we suddenly became office mates in our cozy home office, and when I discussed every minor detail of my PhD project with you. Because of you, this thesis was written in many beautiful places such as sunny Barcelona, cultural Vienna, a desolated German castle (with a grand piano to play on) and an Austrian lake house. I love all of our passions we enjoy together: singing, cooking, visiting theatre, hiking in the mountains, playing games, drinking many cups of tea, traveling across the world... I look forward to all of our upcoming journeys and adventures. I cannot wait!

Finally, I would like to thank all the other people I did not mention in person but with whom I enjoy life every day: *all dedicated UU colleagues, my fellow actors, fellow singers and fellow dancers, my great friends* and *lovely family*. I am grateful!

Hora est!

Marjolein