Promoting Interactivity and Engagement in Tertiary STEM Education using Technology

---

Dissertation
an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig-Maximilians-Unversität München

eingereicht von
Sebastian Mader

07. September 2020

Promoting Interactivity and Engagement in Tertiary STEM Education using Technology

---

Dissertation
an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig-Maximilians-Unversität München

eingereicht von
Sebastian Mader

07. September 2020

Sebastian Mader

Promoting Interactivity and Engagement in Tertiary STEM Education using Technology

Erstgutachter:	Prof. Dr. François Bry Ludwig-Maximilians-Unversität München
Zweitgutachter:	Prof. Dr. Elvira Popescu University of Craiova
Tag der mündlichen Prüfung:	07. December 2020

Eidesstattliche Versicherung

(Siehe Promotionsordnung vom 12.07.11, § 8, Abs. 2 Pkt. .5.)

Hiermit erkläre ich an Eidesstatt, dass die Dissertation von mir selbstständig, ohne unerlaubte Beihilfe angefertigt ist.

München, den 07. September 2020

Sebastian Mader

Abstract

Tertiary STEM education is characterized by an ever-increasing number of student enrollments coupled with a substantially slower increase of the teaching staff. That imbalance often leaves large lectures as the last resort and makes personal interaction between lecturers and students rarer. Promoting students' engagement in their learning and interactivity during courses becomes increasingly difficult under those circumstances.

Technology provides means for addressing and alleviating those problems: Students can be given more stake in their learning, lecturers can be provided with ways to make their large classes more interactive and engaging and be supported in deciding which students require their personal help. As part of this thesis, the learning and teaching platform Backstage 2 was implemented as such a technology. Backstage 2 consists of two main components: An audience response system and a collaborative annotation system. By combining those two components in different configurations, various technology-enhanced learning and teaching formats addressing the aforementioned problems can be created.

Four learning and teaching formats were conceived (or adapted), implemented, and evaluated as part of this thesis: *Large Class Teaching* uses the collaborative annotation system as a backchannel for students and the audience response system to introduce interactivity. *Phased Classroom Instruction* uses the audience response system in combination with subject- and exercise-specific editors to enable more extensive exercises even in large classes. *Collaborative Peer Review* breaks down traditional peer review into a collaborative activity between all stakeholders of the review using the collaborative annotation system. Finally, in *Bite-sized Learning*, technology guides students through quizzes provided by the audience response system.

The contributions of this thesis are threefold: A component-based approach towards creating learning and teaching formats, exemplary formats created using an audience response system and collaborative annotation system, and evaluations for all formats pointing towards their effectiveness. Furthermore, approaches beyond learning and

teaching formats in the form of gamification and game-based learning are explored in the final part of this thesis.

Zusammenfassung

Der tertiäre Bildungsbereich ist von einer stetig steigenden Zahl von Studenten und einem wesentlich geringeren Anstieg von Lehrpersonal geprägt. Dieses Ungleichgewicht führt dazu, dass große Vorlesungen oft der letzte Ausweg zum Unterrichten bleiben und dass dadurch die persönliche Interaktion zwischen Dozenten und Studierenden immer seltener wird. Diese Umstände machen es schwer, die Beschäftigung von Studierenden mit ihrem Lernen und Interaktivität in großen Kursen zu fördern.

Technologie bietet Möglichkeiten, diese Probleme anzugehen und zu lindern: Studierenden kann mehr Anteil an ihrem eigenen Lernen gegeben werden, Dozenten können Möglichkeiten geboten werden, große Kurse interaktiver zu gestalten und können dabei unterstützt werden, zu entscheiden, welche Studierenden ihre Hilfe benötigen. Im Rahmen dieser Arbeit wurde die Lern- und Lehrplattform Backstage 2 als eine solche Technologie implementiert. Backstage 2 besteht aus zwei Hauptkomponenten: Einem Audience Response System und einem kollaborativen Annotationssystem. Durch die Kombination dieser beiden Komponenten in verschiedenen Konfigurationen können verschiedene technologie-gestützte Lern- und Lehrformate umgesetzt werden, mit welchen die angesprochen Probleme adressiert werden können.

Im Rahmen dieser Arbeit wurden vier Lern- und Lehrformate konzipiert (oder adaptiert), implementiert, und evaluiert: Large Class Teaching benutzt das kollaborative Annotationssystem als Backchannel und das Audience Response System um in großen Kursen Interaktivität zu bieten. Phased Classroom Instruction benutzt das Audience Response System in Verbindung mit fach- oder aufgabenspezifischen Editoren um Studierende auch in großen Kursen umfangreiche Aufgaben lösen lassen zu können. Collaborative Peer Review benutzt das kollaborative Annotationssystem um aus Peer Review eine kollaborative Aktivität zwischen allen Teilhabern des Reviews zu machen. Im letzten Format, Bite-sized Learning, führt Technologie Studierende durch Quizze, die mit Hilfe des Audience Response Systems beantwortet werden.

Die Beiträge dieser Arbeit sind dreierlei Art: Ein komponentenbasierter Ansatz für die Implementierung von Lehr- und Lernformaten, beispielshafte Implementierungen von vier Formaten auf Basis eines Audience Response Systems und eines kollaborativen Annotationssystems, und Evaluationen aller Formate, die darauf hindeuten, dass die Formate ihren Zweck erfüllen. Darüber hinaus werden im letzten Teil der Arbeit noch Gamification und Lernspiele als weitere Möglichkeiten zur Förderung von Interaktivität und Beschäftigung mit dem Lernen diskutiert.

Acknowledgements

A work of this scope is not the work of a single person. Many persons were in some way or the other contributors to this work and I would like to use this opportunity to thank them.

First of all, I would like to thank François Bry who first enabled me to pursue this topic and gave me the freedom I needed to pursue the topic. I am incredibly thankful for his constant support and enthusiasm throughout all stages of my research.

Next, there is Niels Heller, who I want to thank for being always available to talk about research and not-so-research-related things, accompanying me to most conferences, and generally being a fantastic support.

I would also like to thank Martin Josko for his technical support, keeping our servers running, and especially repairing them when we broke them again, and Elke Kroiß, for taking care of all the organizational stuff and keeping the teaching unit running.

Special thanks go to Maximilian Meyer and Anna Maier, for their work on the JavaScript editor which first enabled the evaluations of the format Phased Classroom Instruction, Simon Wanner, for creating and implementing the current design of Backstage 2, Manuel Hartmann, for his work on Reification, Korbinian Staudacher for conceiving and implementing the editors for logical proofs, and Konrad Fischer for developing an editor for hierarchical map quizzes.

Furthermore, I would like to thank all other students who contributed to some extent to Backstage 2: Martin Gross, Christian Mergenthaler, Jakob Fürst, Michael Thanei, Max Schwarzfischer, Bastian Heinzelmann, Konrad Fischer, Korbinian Staudacher, Julian Reff, Nikolai Gruschke, Ahmed Shawky, Ziad Mohammad, Martin Matthias, Cedrik Harrich, Jan Sprinz, Vasil Lazarov, and Xiaojie Shi. It was a pleasure working with each and every one of you.

Two projects would have been impossible without cooperation with people of other disciplines: For the course medicine, Franz Pfister came to me with an initial idea for the course and handcrafted together with Konstantin Dimitriadis and Boj Hoppe all the quizzes of the course. Thank you for the chance of working with you.

For the project on Ancient Egypt, I am incredibly grateful to everyone who contributed to the project which includes Julia Budka, Alexander Schütze, Mona Dietrich, Desiree Breineder, Eva Hemauer, and Katharina Rhymer on the side of the Egyptology who proposed the initial idea, worked together with us on all aspects of the course, and curated the quizzes. On the side of Computer Science, there are François Bry, Niels Heller, Konrad Fischer, Korbinian Staudacher, and Elisabeth Lempa who all worked on different aspects of the project. Furthermore, much appreciation to Beatrice Sax for drawing the illustrations of the various structures of Ancient Egypt and the backgrounds used for Reification in the course on Ancient Egypt.

On the personal side, there are Josip Bratic, Christoph Hepting, Marco Lorenz, and Florian Schnell who never failed to provide me with much-needed distraction. My parents, Roland and Carola, who enabled me to pursue the path I am currently pursuing and supported me throughout all stages. Iris, for always being there for me. Thank you.

Previous Publications

Larger parts of the results reported about in this thesis have been published in the proceedings of international conferences or in journals. The list of these publications is as follows:

The author of this thesis was main contributor in the following publications:

- Sebastian Mader and François Bry. “Blending Classroom, Collaborative, and Individual Learning Using Backstage 2”. In: *8th International Conference in Methodologies and intelligent Systems for Technology Enhanced Learning (MIS4TEL 2018)*. Springer, 2018, pp. 3–11
- Sebastian Mader and François Bry. “Gaming the Lecture Hall: Using Social Gamification to Enhance Student Motivation and Participation”. In: *The Challenges of the Digital Transformation in Education - Proceedings of the 21st International Conference on Interactive Collaborative Learning (ICL2018)*. Springer, 2018, pp. 555–566
- Sebastian Mader and François Bry. “Fun and Engagement in Lecture Halls Through Social Gamification”. In: *International Journal of Engineering Pedagogy* 9.2 (2019), pp. 117–136
- Sebastian Mader and François Bry. “Phased Classroom Instruction: A Case Study on Teaching Programming Languages”. In: *Proceedings of the 10th International Conference on Computer Supported Education*. SciTePress, 2019, pp. 241–251
- Sebastian Mader and François Bry. “Towards an Annotation System for Collaborative Peer Review”. In: *International Conference in Methodologies and Intelligent Systems for Techhnology Enhanced Learning*. Springer, 2019, pp. 1–10
- Sebastian Mader and François Bry. “Audience Response Systems Reimagined”. In: *International Conference on Web-Based Learning (ICWL 2019)*. Springer, 2019, pp. 203–216

- Sebastian Mader, Niels Heller, and François Bry. “Adding Narrative to Gamification and Educational Games with Generic Templates”. In: *Proceedings of the 18th European Conference on e-Learning (ECEL 2019)*. ACPI, 2019, pp. 360–368
- Sebastian Mader and François Bry. “Promoting Active Participation in Large Programming Classes”. In: *Computers Supported Education*. Springer, 2020, to appear

In the following publications, both the author of this thesis and Niels Heller were main contributors with equal contributions:

- Niels Heller, Sebastian Mader, and François Bry. “Backstage: A Versatile Platform Supporting Learning and Teaching Format Composition”. In: *Koli Calling ’18: Proceedings of the 18th Koli Calling International Conference on Computing Education Research*. ACM, 2018
- Niels Heller, Sebastian Mader, and François Bry. “More than the Sum of its Parts: Designing Learning Formats from Core Components”. In: *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*. ACM, 2019, pp. 2473–2476

In the following publication, Korbinian Staudacher was the main contributor:

- Korbinian Staudacher, Sebastian Mader, and François Bry. “Automated Scaffolding and Feedback for Proof Construction: A Case Study”. In: *Proceedings of the 18th European Conference on e-Learning (ECEL 2019)*. ACPI, 2019, pp. 542–550

Chapters 4, 6, 7, and 9 are based on previous publications, but were written from scratch, that is, were revised in respect to the referenced literature and their argumentation. Furthermore, across all those chapters, new evaluations and findings are reported. The beginning of each chapter shortly details the publications it is based on and the added evaluations. Outside of those chapters, when referring to those chapters and their contents, similarities are possible as well.

Furthermore, a work of this scope would not have been possible without other contributors: There are parts of Backstage 2 which were developed by students and most often reported on in their respective bachelor or master thesis. Hence, the parts referring to students’ works can share similarities already reported on in the respective bachelor of master thesis. The thesis only refers to works of students that were personally advised and supervised by the author of this thesis.

Contents

List of Figures

List of Tables

1 Introduction

The "massification" (a term used among other by Hornsby and Osman [HO14]) of higher education that took place during the last decades (and still takes place today) brought more students to higher education [Tro99; Var13; Bat10; Big11; MK10], but did generally not involve a corresponding increase of the number of teaching staff which lead to increasing student-to-teacher ratios [Sch91; Tro99; Bat10; WJ92; Gib92]. Anecdotally, at the author's institution, the student-to-professor ratio grew from 142 students per professor to 212 from 2014 to 2018 [Hel+19]. Further evidence of the massification of higher education is provided by a report of the European Commission (see [Cro+17]), from which Heller [Hel20] infers that across the European Union the number of students grew around four times more than the number of teaching staff in the years from 2000 to 2015. Now that an ever-increasing number of students is supposed to be taught by teaching staff which numbers did not increase accordingly, class sizes had to increase in turn which resulted in mass classes attended by a few hundred to thousand of students [Arv14; WJ92; MK10; Sch91]. The emergence of mass classes poses a variety of challenges for teaching staff and students in higher education.

According to Prince [Pri04], "[a]ctive learning is (...) any instructional method that engages students in the learning process" [Pri04, p. 223] and has been shown to increase students' learning achievements in STEM subjects [Fre+14]. However, mass classes often prevent the use of active learning formats as those heavily rely on the interaction between students and lecturers as well as the interaction amongst students. Such forms of interaction are inhibited or difficult to realize in mass classes [Akb+10; Rat+03; Gle86; Gib92; Cot+08]. Take for example the active learning format *flipped classroom* where the parts usually done in the classroom are swapped with the parts usually done outside the classroom, that is, students learn the subject matter outside the classroom using learning material provided by lecturers and classroom sessions are dedicated to exercises and application of the content [BV+13] under the guidance of a lecturer [PK13]. That guidance often comes in form of scaffolding where lecturers "[control] those elements of the task that are initially beyond the learner's capacity, thus permitting him to concentrate upon and complete only those elements that are within his range of competence" [Woo+76, p. 90] and is often coupled with fading where the support is "gradually diminish[ed] until it is no longer needed" [VM+03, p. 5]. It is obvious that support

of this form is highly individual and exactly the bottleneck that makes the format scale badly to large classes, as there are only so many students that can be supported effectively by a single lecturer.

Hence, when lecturers are supposed to teach something to a large group of students, they often resort to the traditional lecture [HO14; MK10; Arv14; WJ92]. Indeed, lecturing seems to be the most common teaching method in higher education [Fre87; Sta+18; Bli00]. While the traditional lecture itself is not inherently bad, as it is effective when it comes to conveying knowledge [Bli00] and represents an economic approach to cope with an increasing number of students [Sch91; Gle86], the traditional lecture is less suited for promoting thought [Bli00]. Another downside is that lectures promote passivity among students [Big11; Fre87] what is associated with decreasing attention: Students' attention is lost after about 10 to 15 minutes of passive listening [Big11]. Furthermore, neither lecturers nor students receive much feedback in large lectures [Big11; Sar12; Gib92; Cot+08; Bli00]: Students refrain from asking questions [Rat+03; Ges92; WJ92] and questions asked by lecturers might only be answered by those students who knew the answer anyway which prevents lecturers from correctly assessing the understanding of their audience [Mar07]. Vice versa, students are getting the impression that everybody around them understood the lecture session's contents as there are no questions, and questions asked by lecturers are answered correctly by their peers.

While receiving feedback is an important aspect during lecture sessions, receiving feedback is generally an important aspect throughout the whole learning process, as feedback ranks among the best teaching methods to promote students' achievement [Hat09]. According to Hattie and Timperley [HT07], feedback "needs to be clear, purposeful, meaningful, and compatible with students' prior knowledge" [HT07, p. 104]. Such feedback is often referred to as *formative feedback* which is feedback that "aims to improve learning while it is happening" [Top+00, p. 150], as opposed to *summative feedback* which measures learning after it has (supposedly) taken place, for example, in form of grades [Top+00]. It is evident that summative feedback is not able to conform to Hattie and Timperley's [HT07] requirements (and its effects, while still positive, are indeed worse than those of formative feedback [Top98]), but giving a few hundreds of students formative feedback in face of limited numbers of teaching staff is hardly realizable [Nic+14; Gib92]. One way to address that issue is through peer review [Nic10], which is "an arrangement in which individuals consider the amount, level, value, worth, quality, or success of the products or outcomes of learning of peers of similar status" [Top98, p. 250]. In short, peer review is a process in which peers provide feedback to their peers. However, organizing peer review for a large number of students only shifts the work from reviewing to organizing which might deter staff from using it.

According to Sarkar [Sar12], "[i]t is these contexts [large classes] that provide useful opportunities for educational technologies" [Sar12, p. 36]. Indeed, these issues can – and have been – addressed using technology: Students have been given a voice in large lectures through backchannels (see, e.g., MiRA [Akb+10], ActiveClass [Rat+03], or the previous version of Backstage [Poh15]), interactivity has been introduced with audience response systems (see, e.g., [BA+13; DB04]), computers can provide automatic feedback to students in many STEM subjects, and technology can provide an environment for and orchestrate the process of peer review (see [LR09] for a review of peer review systems). However, technology is not the panacea for all issues higher education is facing (and it is very unlikely for such to exist) and should not be seen as one: Draper and Brown [DB04] argue that educational technology should not be used for the sake of the technology, but to solve an existing problem, that is, education should always come first.

As part of this thesis, four technology-enhanced learning and teaching formats (called learning formats from here on), that do exactly that – using technology not for technology's sake, but to solve problems of higher education – have been conceived (or adapted), implemented, and evaluated. A learning format is "the 'long term' (...) organization of teaching methods within a course" [Hel20, p. 19] with teaching methods being "a set of principles, procedures, or strategies to be implemented by teachers to achieve the desired learning in students" [Wes08, p. v]. In short, a learning format describes the way a course, or parts of a course, are taught. The learning formats outlined in the following explicitly do not aim at replacing lecturers with technology but to support lecturers and students alike with technology to enable learning and teaching that would otherwise not be possible with a large number of students and a limited number of teaching staff.

The format *Large Class Teaching* was actually conceived by Alexander Pohl as part of his doctoral thesis [Poh15] for the previous version of Backstage. It addresses the lack of feedback and the passivity and anonymity among students in large lectures through technology: A backchannel allows students to communicate anonymously during lecture sessions, and quizzes conducted with an audience response system provide regular breaks that restore students' attention.

Phased Classroom Instruction uses technology to make an active learning format akin to flipped classrooms possible with larger audiences: A lecture session starts with a mini-lecture after which students work alone or in teams on an exercise. Students work on the exercises using problem- or subject-specific editors which provide students with immediate feedback and scaffolding. With the editors supporting students, lecturers have more time at hand to focus on those students for who the scaffolding and feedback provided by the editors is insufficient and require a lecturer's personal support. To identify whom to support, lecturers are supported by

technology which provides them an overview of students' progress on the exercise and suggestions which students most likely require help.

Collaborative Peer Review is a format in which students review their peers' work in a collaborative process. Technology provides students an environment in which they review their peers' works where reviews are shared immediately with the other stakeholders of the review, that is, possible other reviewers and the creator of the reviewed work, who then can react to reviews. By that, possible misunderstandings or unclear reviews can already be addressed during the review phase, and reviewers creating the same review twice is prevented.

Finally, *Bite-sized Learning* is the odd one out of the learning formats, as in this format, the lecturer's only task is to provide learning material while in the other formats, lecturers still had a more prominent role. Lecturers provide quizzes of various types which then can be worked on by students at their own pace while they are provided with immediate feedback on correctness and explanations to the quiz.

The learning formats are part of the learning and teaching platform Backstage 2 which was built from scratch as part of this thesis. Backstage 2 encompasses two main components: A collaborative annotation system and an audience response system. With the collaborative annotation system, students and lecturers alike can annotate lecture material where annotations are shared immediately upon creation with all other users who then can react to them. With the audience response system, lecturers can run quizzes of various types during lecture sessions where each student gives an individual answer using their personal device. These two components have been designed with versatility in mind which made it possible to implement all of the learning formats by combining them in different configurations.

Note that the term Backstage is used for two other projects as well: The original version of Backstage – the foundation of Backstage 2 –, which was conceived, implemented, and evaluated by Alexander Pohl in his doctoral thesis [Poh15], and Backstage 2 / Projects which was conceived and implemented at the same time as Backstage 2 by Niels Heller as part of his doctoral thesis [Hel20]. Hence, to avoid confusion, the term *Backstage 2* is used throughout the thesis to refer to the platform conceived and implemented as part of this thesis, while Backstage and Backstage 2 / Projects are used when referring to the other projects.

Besides learning formats, gamification and educational games have been explored as further avenues for introducing interactivity and engagement to mass classes. One approach consisted of outfitting the audience response system with a gamification based on teams: Each student is part of a team and contributes to their team's score

by participating in quizzes. Furthermore, a generic gamification called *Reification*, and a generic educational game, *Synapses*, were conceived. Both approaches are generic with respect to their narrative, that is, the approaches only provide a frame which can be filled with a narrative that fits the context they are deployed in. However, as both concepts are only partially implemented, no evaluations were conducted.

The contributions of this thesis are as follows:

- The conception and report on the implementation and evaluation of the learning and teaching platform Backstage 2 which includes the implementation of a collaborative annotation system and an audience response system. Backstage 2 aims at being a technological foundation for interactive and engaging learning formats.
- The conception (or adaptation) and implementation of four technology-enhanced learning and teaching formats using Backstage 2's collaborative annotation system and audience response system and report on evaluations of the formats in real teaching contexts in, taken together, 18 courses where the formats were met consistently with positive students' attitudes.
- The conception and report on the implementation and evaluation of a gamification mechanism based on teams in three courses from which conclusions on the applicability of the gamification could be drawn. Furthermore, report on concepts of a generic gamification mechanism and a generic educational game.

This thesis consists of three parts: Part I introduces the learning and teaching platform Backstage 2, its basic structure and features, and its main components, the collaborative annotation system and the audience response system. Part II dedicates one chapter to each of the four technology-enhanced learning and teaching formats. Each of the chapters discusses the motivations for the format, the format itself, and then presents and discusses the results from evaluations of the format. Finally, Part III first gives an outlook at other means for promoting interactivity and engagement in form of gamification and educational games before the final chapter summarizes the thesis and gives perspectives for future work.

Part I

Meeting the Cast

Backstage 2 is a learning and teaching platform built with a component-based architecture in mind: There are two main components (the main actors), the collaborative annotation system and the audience response system, which together with the basic structures and features of Backstage 2 (their supporting cast), can be combined to constitute a variety of learning and teaching formats.

This part first introduces the basic structures and features of Backstage 2, before first the collaborative annotation system and then the audience response system are introduced.

2 Basic Concepts of Backstage 2

Generally speaking, Backstage 2 is a web-based educational software for supporting courses. A course brings together learning material, lecturers, and students to work towards – and ideally to achieve – a common learning goal using the learning and teaching formats described in Part II. While the collaborative annotation system and the audience response systems are the main components of said formats, without the basic concepts and features described in this chapter, they could not be combined into learning and teaching formats.

While in traditional teaching, *course* is often equated with weekly lecture sessions, courses in Backstage 2 are something different: They can be the technological counterpart to a traditional course with weekly lecture sessions, but also provide asynchronous learning activities to students, be completely self-paced without any face-to-face activities, or be a mix of the mentioned aspects. Note that this list is non-exhaustive, as Backstage 2 makes no assumptions on its use; how it is used lies completely in the hand of its users.

The learning material in Backstage 2 comes in the form of *units* which are the building blocks for any learning that takes place on Backstage 2.

2.1 Units

There are two forms of units: *Simple Units* and *Compound Units*. Simple Units are the smallest learning objects in Backstage 2, such as a single page of a PDF document, an image, or a video. Compound Units are a collection of Simple Units of arbitrary type, that is, Backstage 2 supports learning material that comprises of various types of media. For example, a page of a PDF document can be followed by a video, which can be followed by code that can be executed directly from the browser. Figure 2.1 illustrates the connection between Simple and Compound Units.

Compound Units are not a list of Simple Units but form a directed acyclic graph where each node represents a unit. By using a directed acyclic graph, Compound Units can contain branches and so provide more than one way to navigate through a Compound Unit. Figure 2.2 shows an example for a branching Compound Unit.

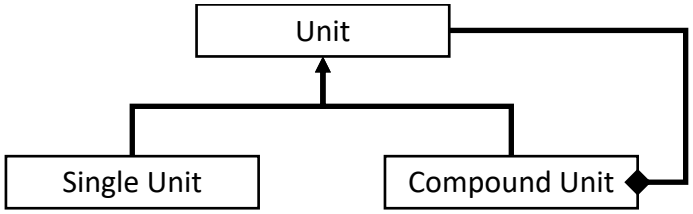


Fig. 2.1.: Two types of units in Backstage 2: Compound Units are a collection of Simple Units.

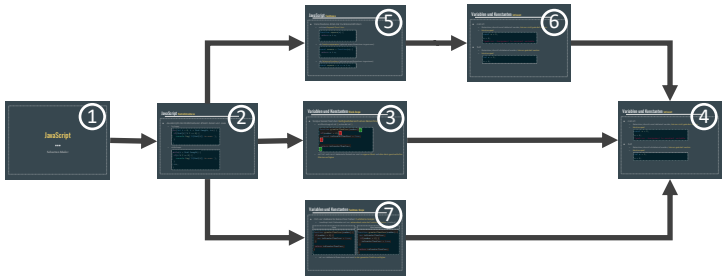


Fig. 2.2.: Example for a branching Compound Unit.

In the example, there is the option to work through the material without detours, that is, nodes 1 to 4, but there is also the option to branch away from the default path at the second unit. The concept of Compound Units was further fleshed out by Max Schwarzfischer [Sch17] in his master thesis: Generally, each Simple Unit can have up to three outgoing paths, a *default* path (nodes 1 to 4 in the example), an *upwards* path (nodes 5 and 6 in the example), and a *downwards* path (node 7 in the example). A possible metaphor for the paths is that an upwards path leads to a more accessible (i.e., *shallow*) representation of the subject matter, while a downwards path is associated with a more sophisticated (i.e., *deeper*) representation of the subject matter. Note that Compound Units using a directed acyclic graph are only implemented in small parts; all evaluations were made with Compound Units consisting of only a default path.

Among the other types of units implemented are *code units* which consist of Markdown (see [Joh04]) interleaved with code editors that already contain code determined by the creator of the unit. Using these code editors, the contained code can be immediately run from the units. An example of a code unit can be seen in Figure 2.2 where two code editors for the programming language JavaScript are interleaved with Markdown. In that way, code units provide interactivity themselves as students can modify the contained code, run it, and observe the output without having to leave the learning material. Lecturers can use the editors, for example, to demonstrate results of code changes. In the figure, an error was included intentionally in the code of the first editor to demonstrate how to fix it during the lecture session.

Fig. 2.3.: Example for a code unit with contains text and program code that can be executed directly from the unit.

At this point, units are not associated with a course and exist independently from courses which makes the same unit reusable in different courses. The next section introduces courses in general and outlines how units can be organized in courses.

2.2 Courses

As already mentioned, courses consist of lecturers, participants, and learning material in the form of units. Units are not put directly into courses but are put into folders that are associated with a course. These folders contain either an arbitrary number of units or an arbitrary number of folders, which again, contain either units or folders. An example of the organization of units into folders can be seen in Figure 2.4. The design of the interface components described and shown in the remainder of this chapter was conceived and implemented by Simon Wanner [Wan17] as part of his master thesis.

Folders are shown in dark blue and can be maximized and minimized by clicking on their title. A minimized folder is just a rectangle, while a maximized folder is shown with an area below where the units or folders contained are shown. In the example, the first two folders are maximized, while the remaining ones are minimized. Both of the maximized folders contain units which are shown in a brighter shade of blue with a title chosen by the lecturer.

The folders associated with a course are displayed on a course's entry page which additionally contains the course's title and description, and optionally, a dashboard.

Fig. 2.4.: Example for structuring units into folders.

Position	Team name	Score
1	Team Blue	4146
2	Team Yellow	3516
3	Team Green	3174
4	Team Red	2733

TEAMS

Fig. 2.5.: Example for a widget in the dashboard which shows the current teams' scores for the social gamification based on teams.

The dashboard consists of individual widgets which display various information about the course. In his master thesis, Wanner [Wan17] implemented various widgets for the dashboard, such as a widget that shows a user's current knowledge of the topics of the course, an overview of a user's current tasks, or an overview of course-related events. However, as Wanner's master thesis focused exclusively on the design part, these widgets were never filled with real content, as no ways of obtaining the data to display were implemented. An example of a widget that was actually filled with real content was developed by the author of this thesis based on Wanner's design and can be seen in Figure 2.5. That widget was used to accompany the social gamification based on teams (see Chapter 9) and displays the current team standings. Except for that widget and a short intermezzo of a widget showing annotation activity, the dashboard was not further utilized.

2.3 Detail View of Compound Units

Upon clicking on a unit in one of the folders (recall, a Compound Unit consisting of several Simple Units), a detail view of that unit is shown. An example for that view can be seen in Figure 2.6. In that view, users can browse through the individual Simple Units of the Compound Unit using the pagination at the top. Below the pagination, the currently selected Simple Unit is displayed which, in the example, is a page of a PDF document. To support Compound Units in form of the aforementioned

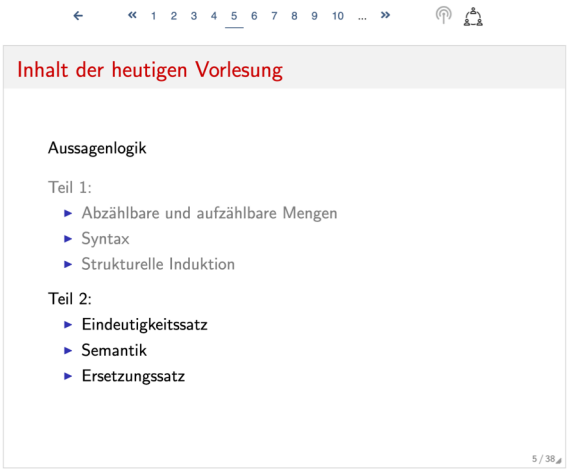


Fig. 2.6.: Detail view of an unit (slide is from François Bry's lecture *Aussagenlogik – Teil 2* licensed under CC BY-NC-SA).

directed acyclic graph, that pagination component would have to be revised to allow navigation between the various paths of a Compound Unit.

Clicking on the icon depicting a radio tower right of the pagination enables automatic synchronization with the lecturer: After clicking that button, the user's current unit follows the navigation of the lecturer, that is, each time the lecturer changes to a unit, the user's unit is changed to that unit as well. That feature was already a feature of the previous version of Backstage. Clicking the icon next to the radio tower reveals (if teams are enabled by the lecturer) a list of teams available to join. Teams are used for the format Phased Classroom Instruction (see Chapter 6) and the social gamification based on teams (see Chapter 9).

Units themselves are mostly static and allow for minimal interaction, but coupling them with the collaborative annotation system and the audience response system makes them the pivotal point for interactions on Backstage 2: Using the collaborative annotation system, every Simple Unit can be annotated and annotations are immediately shared with all other participants. The audience response system allows to attach quizzes of various types to Simple Units which then can be run either during lecture sessions or be done asynchronously by students at their own pace.

2.4 Wrapping up Basic Components

This chapter outlined the various features of Backstage 2 that act as the glue between the collaborative annotation system and audience response system and only these enable those two components to be combined into learning and teaching formats. As mentioned throughout this chapter, the concept for Backstage 2 is much bigger

than what was actually implemented which is natural for software developed as part of a doctoral thesis (actually, any software at all) where it is important to focus on those parts important for research. The following shortly outlines features that were not implemented but would make sense for software similar to Backstage 2.

The dashboard envisioned as part of Simon Wanner's [Wan17] master thesis was with few exceptions never used. As already mentioned, the reason for that is that the functionality which would provide the information to fill the dashboard's widgets, such as assessing a user's level of knowledge or determining what events happened since a user's last login, was never implemented. Nonetheless, something akin to a dashboard is important, as dashboards can provide an "at a glance" [Few06, p. 27] overview of important aspects and can (if built correctly) improve users' awareness [Few06].

Another feature is a task assignment system which can assign tasks to users and check for their completion. This very check for completeness is the crux of the matter, as, in face of mass classes, that check should be done by software. While it might be easy to check whether a user has uploaded or created a document, the more interesting scenarios are more complex as well: Imagine a task where users are supposed to read a scientific article (how does software determine whether a user had read an article?) or review peers' essays using annotations (how many annotations constitute a complete review?).

On the side of lecturers, there is much room for improvement as well: While there are some user interfaces for creating courses, adding material, and so on (which are omitted in this thesis), not everything can be done from the software itself. Many changes have to be done directly in the database. Needless to say, this is in no way acceptable for applications that are used outside of research.

As already mentioned, with few exceptions, the audience response system and the collaborative annotation system are the main drivers of interactivity and engagement in the learning and teaching formats introduced in Part II. The next two chapters introduce first the collaborative annotation system and then the audience response system.

3 Collaborative Annotation System

The first main component from which the learning and teaching formats introduced in Part II are built is a collaborative annotation system. A collaborative annotation system is software that allows users to create annotations to documents of various kinds of media. Annotations are shared with all or a group of users and can often be commented on (see, e.g., VPen [Hwa+11], HyLighter [LL05], and CoNote [DH95]). Among the cited benefits of collaborative annotation is that others' annotations expose one to different ideas and views [Glo+04]; similarly, Su et al. [Su+10] argue that in a collaborative annotation system "learners can collaboratively explore and exploit valuable knowledge" [Su+10, p. 753].

Learners see value in others' annotations: In her bookstore study, Marshall [Mar97] observed that there were students who were explicitly looking for used books containing annotations from previous owners. However, Marshall observed that not all annotations were of the same value to them, as they looked for annotations which were formulated in a way so that another person could make sense of them. However, even though there seem to be annotations which are valuable for other persons besides the creator, sharing them is not an easy task: Either the annotated document itself has to be given away or copies of the annotated medium have to be made [Hof+09]. The issue of sharing can be addressed with digital annotation, and making digital annotation collaborative opens up new ways of interaction not possible when annotating physically, such as commenting on others' annotations [Hof+09].

Collaborative annotation systems have been deployed in a variety of contexts, such as a backchannel (see the previous version of Backstage [Poh15]), for collaboratively creating knowledge to documents such as scientific articles or course material (see, e.g., [Su+10; Raz+12; DH95]), assessing others' work (see, e.g., [Hwa+08]), or submitting homework in form of annotations (see, e.g., [Hwa+11]). This very versatility of collaborative annotation systems is what makes them a sound choice for a communication and collaboration medium of learning and teaching platforms, and hence, Backstage 2 uses a collaborative annotation system for exactly those purposes.

The backbone of Backstage 2's collaborative annotation system is *Annoto*, a framework for implementing so-called *Annotators* which are software components that allow the annotation of a certain kind of media, such as PDF documents, images, or videos, which was conceived and implemented as part of the author's master thesis [Mad15]. Accordingly, even though Backstage 2's collaborative annotation system extends upon the features of Annoto, the main principles are similar, and hence, the referenced literature and argumentation in this chapter are in parts similar to those found in the author's master thesis.

Backstage 2's collaborative annotation system allows users to annotate the units of a course. Annotations can either be private or are upon creation immediately shared with all other participants of a course. Users can interact with annotations by commenting on and up- or downvoting them. As more users who create annotations lead to a greater number of annotations compared to individual annotation, units might become cluttered with annotations. For this reason, Backstage 2's collaborative annotation system provides various means for countering this annotation overload through, for example, means for filtering and searching annotations.

The following chapter gives an overview of the implementation of the collaborative annotation system and its features. Before diving into the implementation, first a short overview of notes, annotations, and their effects as well as other collaborative annotation systems is given. This chapter concludes with an outlook on possible extensions to the collaborative annotation system and future research avenues.

3.1 Annotations and Collaborative Annotation Systems

Notes A concept related to annotations are notes. Annis and Davis' [AD75] description, that students "report for class carrying a notebook in which to take notes on the material presented" [AD75, p. 44] suggests that notes are written commentary detached from the material they refer to. The first research on notes was done by Crawford who found that taking notes can have positive effects on students' learning achievements [Cra25b; Cra25a].

Later research focussed on the *why* as well, that is, what function of notes leads to them improving learning outcomes. According to di Vesta and Gray [DVG72], taking notes serves two functions: An *external storage* function, for which they cite Miller et al. [Mil+60], who suggest that notes act as resources for later review, and an *encoding* function which suggests that by taking notes the content is transformed to a representation that aligns with the note-taker's cognitive structures. Research

on which of both functions note-taking serves is inconclusive: di Vesta and Gray [DVG72] found only evidence for the encoding function, while Carter and van Matre [CVM75] found only evidence for the external storage function. Fisher and Harris [FH73] found evidence for both functions, with external storage being the more important function. On the other hand, Annis and Davis [FH73] found evidence for both functions as well, but found encoding to be the more important function. Regardless of the exact function, research mostly agrees that taking notes has a positive effect on learning. As this chapter is on digital annotation, a more detailed overview of traditional note-taking is beyond the scope of this chapter. Refer to Carrier and Titus [CT79] for a more complete overview of traditional note-taking.

Annotations At the beginning of the author's master thesis [Mad15], a model for annotations is synthesized from the definitions of *note* and *annotation* as found in the Oxford Advanced Learner's Dictionary which finally arrives at that "[a]n annotation consists of a note, i.e., the content part, and a part of the medium the note refers to, which is in the following called context of an annotation" [Mad15, p. 1]. Hence, according to this model, an annotation consists of *content* and *context*.

Using pen and paper for annotation restricts the possible types of contents to what is possible with a pen, while digital annotation introduces the option for other types of contents, such as multimedia content, which Hwang et al. [Hwa+11] state is something that should be supported by web-based annotation systems. Several annotation systems implement contents that would not be possible with pen-and-paper annotation, such as VPen, which supports audio, images, and video [Hwa+11], and HyLighter, which supports audio and video annotations [LL05]. Note that the contents of annotations are independent of the type of media that is being annotated.

Context is what separates notes from annotations: As alluded at the beginning of this section, a note is detached from the material it refers to, while the material an annotation refers to is an integral part of it. Without context, most annotations would be incomprehensible with Hoff et al. [Hof+09] suggesting that in their context-based nature lies "precisely the power of annotations" [Hof+09, p. 222]. Various annotation representation frameworks include means for representing context, such as Annotea through its *context* attribute [KK01] or the Web Annotation Data Model through its *target* attribute [San+17].

In contrast to content, context is dependent on the type of media being annotated: In the author's master thesis, three dimensions for context were identified: There is the *spatial* dimension, which is, for example, used when annotating PDF documents where annotations can refer to regions which can be described using coordinates.

Audio files are an example for the *temporal* dimension: Here, annotations can refer to points in time or time interval. Finally, video offers both contexts in form of the *spatio-temporal* dimension: An annotation can not only refer to regions on the video which can be described by coordinates, but these coordinates may only refer to the content shown at a certain point in time or for a certain time interval. Note that first supporting annotation of multimedia documents creates the need for the latter two types of context, as annotation using pen and paper only uses the spatial dimension.

Refer to the author's master thesis [Mad15] for a more detailed discussion of annotations as in this thesis only those parts required for the understanding of the remainder of the thesis were introduced. Next, a selection of collaborative annotation systems and the results of their evaluations are introduced.

Collaborative Annotation Systems CoNote by Davis and Huttenlocher [DH95] is among the first collaborative annotation systems and supports the annotation of text and HTML documents, as well as commenting on annotations. Annotations cannot be placed at arbitrary positions but only at so-called *annotation points* defined by the author of a document. The authors report anecdotal evidence which suggests that the use of CoNote led to fewer students getting bad grades and that students reported that seeing their peers' annotations made them notice that others were having problems as well.

A later system, EDUCOSM by Nokelainen et al. [Nok+05], supports the annotation of HTML where annotations can be placed at arbitrary positions on documents but cannot be commented on. This is by design, as the authors intend for discussions to take place in document-specific newsgroups. In their evaluation, students had a positive attitude towards EDUCOSM: They thought that EDUCOSM enriched their learning process and led to better studying habits. However, students found their peers' highlights (i.e., only a context without content) annoying but thought that their peers' comments (i.e., annotations that include content) to help their learning.

Another system that emerged around the same time is CASE by Glover et al. [Glo+04] which supports the annotation of HTML documents. Annotations can be shared with other users but seemingly not be commented on. HyLighter by Lebow and Lick [LL05] emerged around the same time and supports annotations that refer to passages in HTML documents. Annotations can have text, graphics, or audio as content and can be commented on. HyLighter offers a unique feature which collates the annotated passages: Passages only annotated by the current user, passages, annotated by other users but not the current user, and passages annotated by other

users as well as the current user are all shown in different colors. For the latter two, a more intensive shade of the color is used the more users annotated the respective passage. The authors report on a field test which showed that HyLighter had positive effects on "participation, engagement, [and] accountability" [LL05, p. 4] and can also "increase the productivity of document-centred group work" [LL05, p. 4]. In another evaluation of HyLighter by Razon et al. [Raz+12], students showed a positive attitude towards HyLighter, but even though a group using HyLighter consistently showed higher learning achievements than a group working with paper copies, the differences between the groups were not significant.

VPen by Hwang et al. [Hwa+07] supports the annotation of HTML with text, pictures, and audio as possible contents. In the version described in the article, each user annotates an individual copy of the document and other students have the option to view other students' documents and annotations, but cannot comment on them or create own annotations on those documents. In an evaluation, the authors compared three usage scenarios of VPen, individual annotation, having access to annotations of a group, and having access to all annotations to individual reading. Students generally had a positive attitude towards VPen and students of the group using VPen showed significantly higher learning achievements than students of the group who engaged in individual reading across all usage scenarios. However, these learning achievements did not translate to higher examination results where no significant differences between the groups were found. The authors suggest that this might be due to all students being motivated to score high in the examination, but speculate that students who did individual reading had more "catch-up work" [Hwa+07, p. 697] to do. Another evaluation of VPen by Hwang et al. [Hwa+11] came to the result that students rarely benefit from viewing their peers' annotations. Subsequent interviews with students suggested that students had problems making sense of their peers' annotations.

The collaborative annotation system PAMS 2.0 by Su et al. [Su+10] supports among other the annotation of PDF and HTML documents with either freeform figures or highlights with textual content. In their study, they compared collaborative group annotation using PAMS 2.0 with group reading using a wiki. Students showed a positive attitude towards PAMS 2.0. In the first round of their study, no significant difference in learning achievement between the groups was found but starting with the second round, the group using PAMS 2.0 showed significantly higher learning achievement than the group using a wiki. The authors suggest that this might be due to students first having to become acquainted with the system before being able to benefit from it. Regarding differences in the examination, the authors found no significant differences between the groups and suggest that students are generally motivated to do well in examinations and that maybe those students who did group reading in a wiki had "more catch-up learning tasks" [Su+10, p. 764] to do. That

result is consistent with the result obtained by Hwang et al. [Hwa+07], who these authors mention as well.

In summary, there are a variety of approaches to collaborative annotation systems in regards to their approach to sharing, commenting, what and what parts of a document can be annotated, and what can be used as content. Regardless of that, students showed positive attitudes towards the use of collaborative annotation systems across all studies, and their use is often associated with an increase in learning achievement.

In two studies, other students' annotations were perceived negatively: In the evaluation of CoNote [DH95], students found their peers' annotations without content annoying. A possible explanation for that attitude can be found when looking at Marshall's bookstore study [Mar97] where students looked for used books that were annotated in a way that allowed them to make sense of the annotations. However, an annotation without content can rarely make sense to others and might be perceived as only cluttering the document, which would explain the negative attitude towards them. In an evaluation of VPen [Hwa+11], students stated that they could rarely make sense of their peers' annotations. This might be due to the system being evaluated in secondary education (whereas all other studies described in this section were at least done in tertiary education) where students might have less developed annotations practices which could make their annotations less valuable to their peers.

3.2 Backstage 2's Collaborative Annotation System

This section introduces Backstage 2's collaborative annotation system and how it can be configured so that it can be deployed in various contexts. The collaborative annotation system allows participants of a course to create annotations referring to any unit of that course which are immediately shared with all other participants. Participants can react on annotations either through commenting or voting on them. The collaborative annotation system extends upon the backchannel of the previous version of Backstage as described by Pohl [Poh15], and hence, borrows concepts from the previous version, such as the available voting options, some of the available purposes for annotations, and the three-step process for creating an annotation which is detailed later in this section. The design of the interface components introduced in the following section was conceived and implemented by Simon Wanner as part of his master thesis [Wan17].

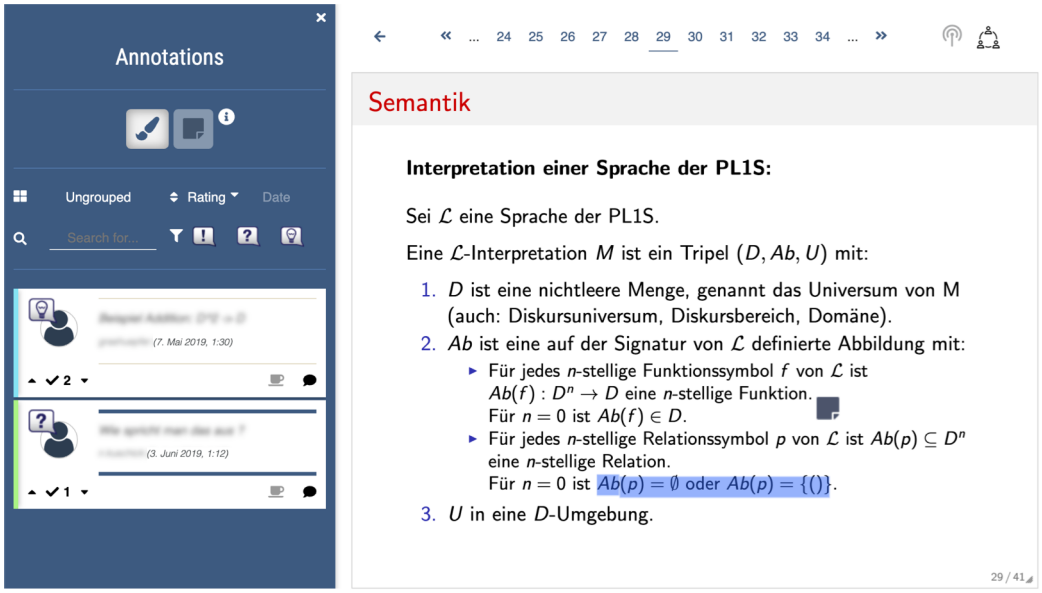


Fig. 3.1.: Detail view of a unit with annotations and an unfolded annotation sidebar (slide is from François Bry's lecture *Prädikatenlogik – Teil 1* licensed under CC BY-NC-SA).

3.2.1 Annotations in Detail Views of Units

The collaborative annotation system exists within the detail view of units which was already introduced in the previous chapter (see Section 2.3). While in the previous chapter, the annotation functionality was omitted, Figure 3.1 now shows an example for that view with annotation functionality shown.

The unit in the example is a page of a PDF document for which two annotations have been created. The *contexts* of these annotations are shown directly on the unit on the right side of the figure: One annotation refers to a single position on the unit which is indicated by the grey icon depicting a note, and the second one refers to a passage of text which is indicated by the blue rectangle enclosing a passage of text. The corresponding *contents* of the annotations are shown in the lower part of the sidebar on the left. For the sake of simplicity, these contents shown in the sidebar are referred to as annotations in the following. Clicking on either content or context highlights the other part. In the example, the second annotation in the list is selected and hence, the respective context is shown in blue on the unit. An unselected context would be displayed as a yellow rectangle.

The sidebar consists of three parts: The buttons at the top allow to change between the contexts which are available for the type of the current unit. Below that, options for filtering, grouping, searching, and ordering annotations are arranged. The remainder of the sidebar consists of a list of annotations that refer to the current unit. The sidebar can be minimized, whereupon it becomes a small stripe at the side

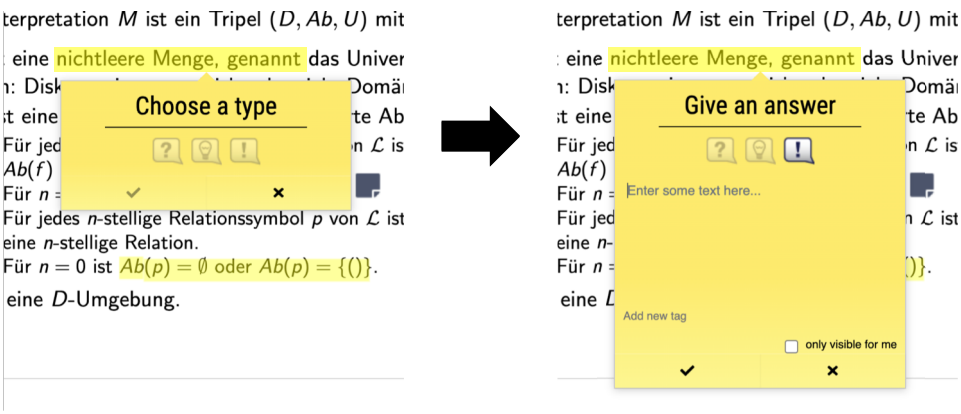


Fig. 3.2.: Process for creating annotations: After selecting a context, a purpose has to be selected before the content of the annotation can be input (slide is from François Bry's lecture *Prädikatenlogik – Teil 1* licensed under CC BY-NC-SA).

not showing any content. Before discussing the representation of annotations in the sidebar in more detail, the process of creating annotations is outlined.

3.2.2 Creating Annotations

For PDF units, two types of context are available: Passages of text, which are created by dragging the mouse with the left button pressed over the desired passage of text, or single positions on the document, which are created by clicking on the desired position.

After a context has been created, a prompt is shown which leads through the process of creating the remainder of the annotation. The two steps of the process can be seen in Figure 3.2: First, the purpose of the annotation has to be selected. There is no purpose selected by default to force participants to make a conscious decision for an appropriate purpose for their annotations. In the example, three purposes are available: asking a question, adding a remark, and answering a question. Which purposes are available is chosen by lecturers so that the purposes fit the context the collaborative annotation system is used for. After a purpose has been selected, the text area in which the textual content of the annotation can be input becomes visible. Further options in that view are to add tags using the text field directly below the text area and setting the annotation private by checking the checkbox.

The process of creating annotations is an adaption of Pohl's [Poh15] three-step process for creating backchannel posts which encompasses selecting a context, then purpose for the post, and only then being able to enter the content of the post. Pohl lists three reasons for that design: first, to increase the effort of creating a post so that student are deterred from creating irrelevant posts (citing the messaging threshold theory [Rei+96]), second, to get students to think about their post through

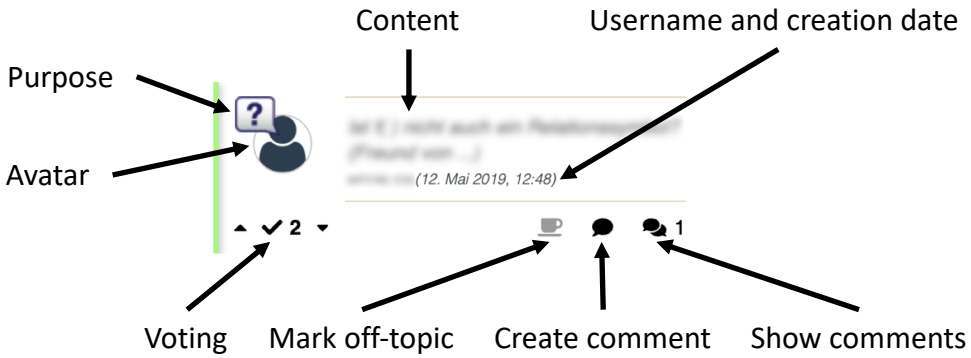


Fig. 3.3.: Representation of an annotation in the sidebar.

a prolonged process (citing Peters [Pet14]), and finally, to act as a facilitating script that guides students through the process of creating posts (citing various research on scripts, see [Poh15]).

After the content has been input and confirmed by clicking the button labeled with the checkmark icon, the annotation has been successfully created. The annotation is now permanently visible both on the unit in form of its context as well as in the list of annotations in the sidebar. As already mentioned, annotations are immediately synchronized with all other participants, that is, immediately after creation, an annotation becomes visible to all other participants without the need of reloading Backstage 2.

3.2.3 Interacting with Annotations

Coming back to the representation of annotations in the sidebar which shows the contents of annotations and from which the various means of interacting with annotations are available. Figure 3.3 shows a labelled version of an annotation's representation in the sidebar.

In the top left part of an annotation, the avatar of its creator is shown superimposed with an icon that represents the purpose of the annotation. In the example, the user has no avatar and hence, is represented through a generic icon, and the annotation was created to ask a question. Right to the avatar and purpose, the content, the username of the creator, and the creation date are shown. Through the buttons at the bottom, the various means for interacting with annotations become available: With the up- and downward-pointing wedges an annotation can be up- or downvoted (representing agreeing or disagreeing with an annotation). The number between the wedges is calculated as the difference between up- and downvotes or zero if that difference is negative. Hence, a high number represents an annotation well-regarded by other students. With the buttons right to that, an annotation can be marked

Fig. 3.4.: Annotation with unfolded comment.

as off-topic (a feature adapted from the previous version of Backstage), create a comment, or unfold already created comments, respectively. For the annotation in the figure, there already exists a comment which is indicated by the number 1 shown next to the icon.

Figure 3.4 shows the same annotation with unfolded comments: Comments are shown slightly indented below an annotation and are represented in the same way annotations are except that no commenting functionality is available as replies to comments are intended to be created as comments to the annotation. For creating comments, the same process as for creating annotations (see Figure 3.2) is used with the exceptions that no context has to be selected and that the forms for selecting a purpose and entering the content are not shown on the unit, but directly below the annotation in the sidebar.

3.2.4 Countering Annotation Overload

As already mentioned, in a collaborative annotation system not a single user, but an indeterminate number of users create annotations which can lead to units becoming cluttered with annotations. However, these very cluttered units might be of special interest for users (after all, why else would they be so heavily annotated?) but are hard to work with precisely because of the high number of annotations. Hence, an important aspect of collaborative annotation systems are means for countering that annotation overload. Figure 3.5 shows the options for grouping, ordering, searching, and filtering annotations which are available in Backstage 2's collaborative annotation system. The component shown in the figure is arranged above the list of annotations as can be seen in Figure 3.1.

Annotations can be ordered either by rating or creation date and either ascending or descending by clicking on the desired ordering criteria. The rating of annotations is calculated from the number of up- and downvotes in the same way as in the previous version Backstage as the "[l]ower bound of Wilson score confidence interval for a

Fig. 3.5.: Available options for grouping, ordering, searching, and filtering.

Bernoulli parameter" [Eva09] as proposed by Evan Miller [Eva09]. Furthermore, annotations can be searched by entering a term in the text field which searches the textual contents as well as the usernames of the creators of all annotations created for the current unit and only shows those which match the entered term. Annotations can be filtered by purpose as well: Each icon represents a purpose and clicking one of them hides or shows, respectively, all annotations of that purpose.

Finally, annotations can be grouped by either purpose or creator. Grouping puts annotations that have the same value in the chosen grouping category in the same group. For example, when grouping by creator all annotations created by the same user would be put in the same group. Each group is shown as a separate sublist in the annotation sidebar and gets assigned a unique color in which the contexts of annotations in that group are shown on the unit. An example of grouping by purpose can be seen in Figure 3.6: For each purpose, a separate sublist is shown in the sidebar with the color of the header of a sublist being the color assigned to that group. On the unit on the right, the context of each annotation is shown in the same color as the color of the group the annotation is assigned to. Each sublist can be minimized by clicking on the minus sign on the left side of the headers which hides the contexts of the annotations of that group as well.

Another aspect of countering annotation overload are ways to identify unread annotations: When the sidebar is minimized, unread annotations for the current unit are indicated by the sidebar glowing green. When the sidebar is unfolded, unread annotations *beat* twice (i.e., the annotation grows, gets smaller, grows again, gets smaller, and then assumes its actual size) and are then shown in boldface. Both cues work well in combination with the immediate synchronization to draw attention to incoming annotations both in a direct way with an unfolded sidebar as well as in an unobtrusive way with a minimized sidebar.

3.2.5 Extending the Collaborative Annotation System

This section introduced the collaborative annotation system exclusively using examples of annotations referring to PDF units, which are, indeed, the kind of units for

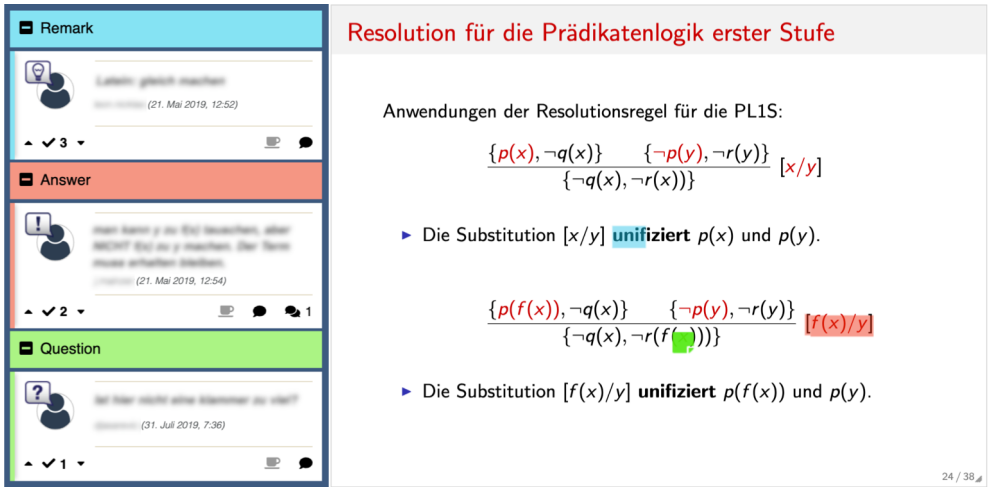


Fig. 3.6.: Example for grouping annotations by type: Each purpose is given an own color and the contexts of annotations of that purpose are colored that way. The annotation list is divided in sublists for the different groups (slide is from François Bry's lecture *Resolution* licensed under CC BY-NC-SA).

which most functionality exists. The reason for that is simple: Most of the existing teaching material has already existed in the form of PDF documents. Besides for PDF units, some annotation functionality is implemented for images and markdown units. When annotating images, rectangles and arbitrary-shaped polygons can be chosen as context; when annotating markdown units, passages of text can be selected as context.

Regarding the content of annotations, only a single type of content was implemented, namely textual content. To add new types of content, a browser-based editor in which content of that type can be created as well as a way to display the output of the editor in a browser have to be implemented (or have to already exist). For example, to support graphical content, a component which allows to draw images would have to be implemented (which then would be shown instead of the text area in Figure 3.2); the drawn image could then be shown using HTML's `img` tag (which then would be shown in the area labeled *Content* in Figure 3.3). Once implemented, a content can be used in combination with all contexts.

Contrary to that, contexts are, as already mentioned, specific to a certain type of media. However, Annoto provides an interface to easily define new contexts: Various events that happen on a unit (e.g., movement of the mouse cursor, pressing a mouse button, ...) can be associated with functions that are executed when the event occurs. For an easy example, there is selecting a passage of text which consists of a function that is executed when the left mouse button is released. This function checks whether some part of the unit is selected, and if so, a context is created out of that selection. Contexts for audio and video units were implemented as part of

the author's master thesis [Mad15] but never made their way into Backstage 2's collaborative annotation system.

The other features of the collaborative annotation system described in this section, that is, the sidebar, the means for filtering, searching, grouping, and ordering, the highlighting of unread annotations, the rating of annotations, the commenting on annotations, and the immediate synchronization of annotations all work independently from the type of unit, context, or content. Hence, new types of media can easily be added with only context(s) for that type of media having to be implemented. Then, already existing contents can be used or – if required – new types of content can be implemented.

3.3 Wrapping up Collaborative Annotation System

This chapter introduced Backstage 2's collaborative annotation system which allows users to annotate units where the annotations are immediately shared with the other participants of a course. The collaborative annotation system is the main component in two of the learning and teaching formats described in Part II: In Large Class Teaching (see Chapter 5) as a backchannel during lecture sessions and for discussing lecture material afterward, and in Collaborative Peer Review (see Chapter 7) for reviewing students' essays. In both formats, the collaborative annotation system is used to annotate PDF units with the main difference being that different purposes for annotations are available.

While there is already a variety of contexts in which the collaborative annotation system in its current form can be deployed in, there are functionalities that would further increase its versatility. These functionalities come either in form as extensions to already implemented features or as completely new features.

Regarding extensions of already implemented features, one of the more obvious features is support for new types of units which would increase the number of contexts in which the collaborative annotation system can be deployed in. Take for example videos which are gaining in importance during the current COVID-19 pandemic where much learning material is made available in the form of videos. Furthermore, new types of content would increase the system's versatility as well: Similar to other collaborative annotation systems (see, e.g., VPen [Hwa+11]), support for audio, graphical, and video annotations could be implemented.

As for new features, providing different levels of visibility of annotations might open up new scenarios for group work. Even though the collaborative annotation system

is already a system for group work when taking the participants of a course as a single group, there are scenarios where working as one big group is not desirable. Take for example a large course where students are supposed to collaboratively make sense of a scientific article using the collaborative annotation system. In this case, having all participants collaborate is likely less effective than collaborating in a small group, because in small group individual contributions are less likely to get drowned out by a large number of annotations. Currently, only two levels of visibility exist: Visible to all participants and visible only to the creator of the annotation. Among the conceivable levels of visibility are visible to a group of participants (see, e.g., PAMS 2.0 [Su+10]), friends, or a user-defined group of participants. In the same vein, giving not only right to view but the right to edit annotations to a group of participants would make it possible to not only collaborate *using* annotations but to collaborate *on* annotations as well.

Another feature would be to support units that consist of more than one type of media such as an image embedded in markdown which otherwise contains only text. Here, for annotating the image, contexts for image units should be available, and for annotating the text, contexts for markdown units should be available.

Finally, while the collaborative annotation system helps users to get an overview of what is unread in the scope of a Simple Unit, for getting an overview of what is unread in the scope of a Compound Unit, the complete unit has to be browsed. Hence, communication awareness (suggested by Hoff et al. [Hof+09] as an important aspect of collaborative annotation systems and already proposed by Pohl et al. [Poh+12] as a possible improvement to the backchannel of the first version of Backstage) should be improved: Users could be notified about comments on their annotations (such as an answer to a question) or new annotations referring to their units (such as a new review for their essay). Nikolai Gruschke developed as part of his (unpublished) bachelor thesis approaches for improving communication awareness, but these approaches never made it into the system long due to bugs (outside of Gruschke's control). These approaches are outlined in more detail in the last part of Chapter 7.

As the collaborative annotation system was exclusively evaluated as part of various learning and teaching formats, future research should focus on an evaluation of the collaborative annotation system outside of such as well. Of interest in such evaluations are user experience and usability, but especially the effectiveness of the proposed means for countering annotation overload.

Next to the collaborative annotation system, there is the audience response system as the other main component of the learning and teaching formats which is introduced in the next chapter.

4 Audience Response System

This chapter is based on the following article:

- Sebastian Mader and François Bry. "Audience Response Systems Reimagined". In: *International Conference on Web-Based Learning (ICWL 2019)*. Springer, 2019, pp. 203–216

In addition to the contents of the article, this chapter includes a description of Backstage 2's audience response system.

After the previous chapter introduced the collaborative annotation system as one of the main components of the learning and teaching formats described in Part II, this chapter now introduces the other main component, the audience response system and how it expands upon what today's audience response systems are offering. Audience response systems are educational technology where first, students respond to quizzes posed by lecturers using technology, and second, the students' responses are aggregated and are then shown to the audience [Cal07]. In that way, audience response systems provide feedback to lecturers which allows them to adapt their teaching [Cal07; KL09; HA13; DB04] as well as to students who are enabled to assess how their understanding relates to their peers [KL09; DB04].

Audience response systems first emerged as clickers [Hun+16], which are small handheld devices equipped with several buttons and a transmission module which allows students to input and transmit their answer to a quiz [Cal07]. Through their limited means of input, clickers support generally only a limited number of question types such as multiple choice or numerical input [She16; MB13; Ima14; GT+13]. Today's omnipresence of smartphones, tablets, and laptops led to the emergence of audience response systems that run on these devices directly from their web browsers [GT+13]. Even though these devices would provide more means for input [HA13; Ima14; GT+13], audience response systems running on these devices are still most often restricted to multiple choice or open answer questions [Bry+14; Sch+15; HA13]. Indeed, a survey conducted by the author of this thesis came to the same result with the majority of examined audience response systems only supporting either multiple choice, open answer, or both [MB19a].

As the effectiveness of multiple choice questions to promote higher-order thinking is debated (see, e.g., [Hal96] for a proponent, and [Whi93; SH12] for opponents),

including question types that go beyond multiple choice into audience response systems might make sense. Learning with questions that promote higher-order thinking is important: A study by Jensen et al. [Jen+14] has shown that students who were tested exclusively with questions promoting higher-order thinking significantly outperformed in an examination students who were tested with recall-oriented questions even in the questions that focussed on recall.

uRespond by Bryfczynski et al. [Bry+14] is an audience response system that supports question types beyond multiple choice so that "more authentic and meaningful questions" [Bry+14, p. 358] can be asked of students. As a reason for the need to go beyond multiple choice, these authors cite a previous study by Cooper et al. [Coo+10] which found that students who were able to select a correct chemical structure from a list of structures (i.e., a multiple choice question) might not necessarily be able to actually construct a chemical structure. Similarly, Hauswirth and Adamoli [HA13] argue that question types that go beyond multiple choice are richer as students "have to 'construct' the solution (...) themselves, instead of picking among a set of solutions" [HA13, p. 500]. Furthermore, they state that such questions leave more room for errors; and errors are something students learn from. Hence, to promote higher-order thinking during classroom sessions, Backstage 2's audience response system supports a variety of problem- or subject-specific editors which are used by students to create their (now more elaborate) answers to quizzes. Adding more types of quizzes was already proposed by Pohl at the end of his thesis [Poh15] where he suggested programming quizzes, that is, quizzes where the students' answers is code, as a conceivable new quiz type.

Furthermore, most audience response systems limit quizzes to two phases: Giving answers and showing aggregated results. Backstage 2's audience response system extends upon that by allowing quizzes to span an arbitrary number of phases, such as a quiz in which students review another student's submission before showing the results.

As quizzes become more complex through the aforementioned problem- or subject-specific editors, it becomes, in turn, more difficult for every student to produce an – even incorrect – answer. Hence, adapting the editors and the quizzes' contents might empower more students to be able to create an answer to a quiz. Such adaptivity was envisioned as the third area in which audience response systems can grow during the creation of Backstage 2's audience response system but was never implemented.

This chapter is structured as follows: First, an overview of research on audience response systems is given. Then the results of the aforementioned study examining the current state of audience response systems conducted by the author are shortly outlined. Following that, the three areas in which audience response systems can

grow are further detailed and illustrated using Backstage 2's audience response system. The last section summarizes the chapter and presents perspectives for future research.

4.1 Audience Response Systems

The use of audience response systems is associated with various positive effects: In their literature review, Kay and LeSage [KL09] identified various benefits associated with the use of clickers. Among these benefits are positive student attitudes towards the use of an audience response system and positive effects on attention, engagement, and generally, interaction during classroom sessions.

Similar observations are made by Cladwell [Cal07] who concludes at the end of her literature review that audience response systems "seem to enhance students' active learning, participation, and enjoyment in classes" [Cal07, p. 19], but to have only "neutral or positive effects (...) on learning outcomes" [Cal07, p. 19]. However, the author reported more positive effects on learning outcomes being observed when audience response systems were combined with peer learning.

A meta-survey conducted by Hunsu et al. [Hun+16] grouped potential outcomes affected by the use of audience response systems into cognitive and non-cognitive outcomes. While the authors found that the use of audience response systems had positive effects on outcomes of both groups, the effects on non-cognitive outcomes were greater than those on cognitive outcomes. Regarding cognitive outcomes, they found a positive effect on knowledge transfer, but no effect on retention. The largest effect reported by the authors on a non-cognitive outcome was found for students' self-efficacy, but positive effects were found, among others, for engagement and participation as well. From the results on cognitive outcomes, the authors conclude that audience response systems might best be utilized to promote higher-order thinking. Another finding of their study was that there was no significant difference between lecture sessions using an audience response system and lecture sessions using quizzes without technological support which indicates that not audience response systems, but quizzes are the important component. Note that this meta-survey considered both clickers and audience response systems running on users' devices, but did not treat them differently.

Doing quizzes without an audience response system might be possible in small courses but becomes more and more difficult with increasing course size: Fear of being incorrect and public humiliation deters a majority of students from answering quizzes [Cal07; Mar07]. Similar problems occur when using show of hands as

quizzing method: The lack of anonymity leads to students either not raising their hand at all or joining the first students who provided an answer [RB06]. Audience response systems provide anonymity to students which makes them a quizzing method that allows every student to answer a quiz without having to worry about giving an incorrect answer [Mar07] which makes them a fitting tool to engage and bring interactivity to today's large class lecture sessions.

4.2 Audience Response Systems Reimagined

In a study conducted by the author [MB19a], a total of 81 audience response systems were identified using a structured approach using Google Search. Afterward, the supported question types of each system were identified by a single judge which yielded the following list of 12 different question types (taken verbatim from [MB19a, p. 207f.]):

- *Choice*: Users select one or more answers from a list of answers.
- *Open answer*: Users enter their own answer.
- *Region*: Users select a point or a region on an image as answer.
- *Sketch*: Users sketch their answer using a drawing tool running in the browser.
- *Fill-in-the-blank*: Users fill blanks in a text, either by entering terms or selecting those from a list of options.
- *Scale*: Users select their answers in a certain range of values using a slider element.
- *Order*: Users arrange a number of items in sequence.
- *Sort*: Users select from pre-defined classes for pre-defined items.
- *Graph*: Users give their answers as a graph created by graphing software running in the browser.
- *Text highlight*: Users select a part of a text as their answer.
- *Match*: Users create pairs from an even number of items.

The percentage of systems implementing a certain question type can be seen in Figure 4.1: Choice and open answer are the only types which were implemented by more than half of the systems; the remaining types were implemented by at best a fourth of systems, but the general trend is that a question type is only implemented by few systems.

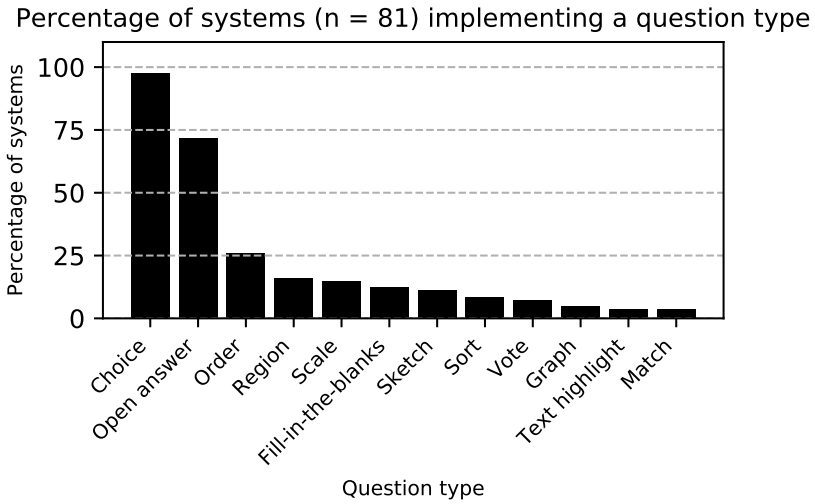


Fig. 4.1.: Percentage of audience response systems implementing a certain question type (adapted from [MB19a, p. 208]).

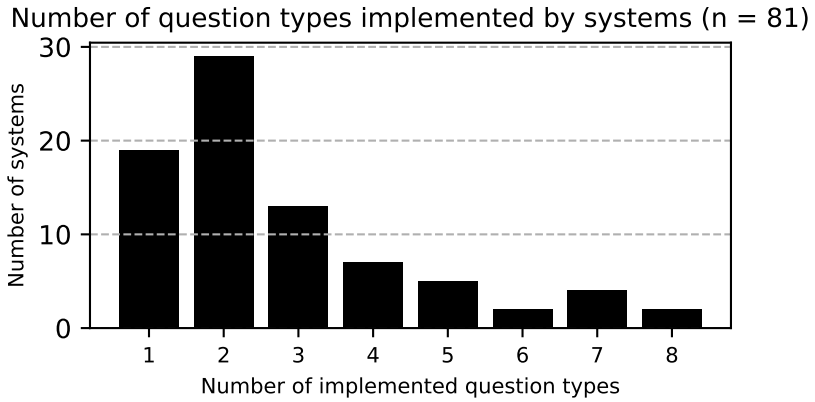


Fig. 4.2.: Number of question types implemented by the examined audience response systems (adapted from [MB19a, p. 209]).

Figure 4.2 shows the number of systems having implemented a certain number of question types. 29 of the 81 examined systems implemented only two question types, and from these systems, 24 implement the combination of choice and open answer. From the 19 systems which implemented only a single question type, 18 implemented choice. Taking these two observations together reveals that over half of the examined systems implemented either only choice or the combination of choice and open answer. With an increasing number of question types, the number of systems supporting that number decreases with only eight systems implementing six or more question types.

There were, however, few audience response systems which provided functionality that went beyond the aforementioned question types:

- Schön et al. [Sch+15] developed an audience response system where question types can be built from a selection of predefined elements which is more flexible than being restricted to predefined question types.
- Informa supports various editors, such as editors for class diagrams, regular expressions, and heap and stack diagrams, in which the answers to quizzes can be created [Hau08; Hau11].
- uRespond includes an editor that allows to construct chemical structural formulas as answers to quizzes [Bry+14]. TopHat[1] provides a similar editor as well as editors that allow giving an answer in the form of a chemical or mathematical equation.
- Two question types were only found in a single system each: *2x2 Pairs* from MentiMeter[2] and *Scrambled Answers* from Quizalize.[3]

There are limitations to the study: Audience response systems were identified from search results on Google Search. Including scientific databases in the review might yield further question types. Furthermore, the identification of question types supported by a system was done by a single judge which could have biased the results.

In summary, the results of the study suggest that today's audience response systems still share many similarities with their hardware predecessors: Even though smartphones, tablets, and laptops would provide more means for inputting an answer (see, e.g., [HA13; Ima14; GT+13]), these means are rarely exploited in today's audience response systems which most often only implemented choice and open answer. This section omitted the description of the methods of the study; refer to [MB19a] for the complete report on the study.

That is not to say that choice and open answer are bad question types, but that restricting to them might not be the best idea as both have disadvantages:

- As already mentioned, the ability of choice questions to promote higher-order thinking is still debated. Furthermore, choice questions leave less room for errors; making errors, however, is important as students learn from making errors [HA13].
- While open answer allows for more complex answers, checking those answers automatically for correctness becomes more difficult the more complex answers become: Checking whether a single term matches the desired answer

[1] `https://tophat.com/`
[2] `https://www.mentimeter.com/`
[3] `https://www.quizalize.com/`

is easy, but as soon as an answer begins to span a whole sentence or more, automatically checking for correctness might no longer be possible. However, only such an automatic check for correctness is what enables immediate feedback and is, therefore, a cornerstone of audience response systems and their question types.

Hence, *reimagining* audience response systems pertains not only creating question types that go beyond choice and open answer but at the same time ensuring that those question types can still be automatically checked for correctness. The next section introduces Backstage 2's approach to question types and an overview of supported question types.

4.2.1 Question Types

One approach for supporting various question types is through subject- and exercise-specific editors, as, for example, done by Informa [Hau08; Hau11; HA13], uRespond [Bry+14], and, to some extent, TopHat. Examples for subject-specific editors are code editors or TopHat's editor for mathematical equations; an example for a problem-specific editor is uRespond's editor for creating chemical structure formulas. Using such editors in an audience response system allows students to create an artifact (their answer) that relates to a subject or a class of problems.

For many subjects and topics, such editors can represent the artifact internally in a way that can automatically be checked for correctness. Furthermore, they can display the artifact and provide available interactions with the artifact in a way that supports students while they work on their answers. Supporting students while they work on their answers is important, as more complex quizzes make differences in knowledge between students more evident: While with choice quizzes, every student can – even if incorrect – easily produce an answer, in more complex quizzes, not all students might be able to produce an answer. Not being able to produce an answer might demotivate students and deter them from participating in future quizzes which is contrary to what audience response systems want to achieve. Hence, as soon as answers created using a problem- or subject-specific editor get more complex, these editors should be built with support in mind so that more students are empowered to produce an answer.

During the work on Backstage 2, a variety of editors for its audience response system have been implemented and found their use in the learning and teaching formats described in Part II. This section shortly introduces the editors and where those editors are used in Backstage 2.

An editor for locating structures on maps of Ancient Egypt was implemented by Konrad Fischer for a course on Ancient Egypt and is described alongside the course in Chapter 8. The same chapter introduces a course on medicine for which an editor for marking symptomatic regions on medical images was built. Note that due to time constraints, that quiz was not built using the audience response system but by misappropriating the collaborative annotation system. With a bit more time, an editor for realizing such question types could have easily been built. An editor for the programming language JavaScript was implemented first by Maximilian Meyer [Mey19] as part of his master thesis and then improved by Anna Maier [Mai19] as part of her master thesis for the format Phased Classroom Instruction and is described in detail in Chapter 6. Similarly, editors for arbitrary programming languages are imaginable but were not implemented as part of this thesis.

Editors for more conservative question types were implemented as well so that Backstage 2's audience response system supports choice, open answer, and fill-in-the-blanks which were implemented by Jacob Fürst as part of his (unpublished) bachelor thesis. Furthermore, an editor for scale quizzes which allows for input of numbers in a certain range was implemented as well.

Editors for Resolution and Natural Deduction Two editors not talked about in the remainder of this thesis are editors for building logical proofs using the proof techniques Natural Deduction and Resolution. Both editors are exercise-specific, as they only allow students to work on exercises where they apply the respective proof technique, but are no general editors for propositional or first-order logic. These editors were conceived, implemented, and evaluated alongside one of the courses described in Chapter 5 by Korbinian Staudacher for his bachelor thesis [Sta18] on which a research paper [Sta+19] is based on. The following is based on [Sta+19] and introduces the editors and explains how the editors support students while they work on exercises. Refer to [Sta18] and [Sta+19] for more detailed descriptions of the editors and results of their evaluation.

Figure 4.3 shows the editor for Resolution. Resolution is a proof technique which can show the unsatisfiability of a set of clauses [Rob65]. In the editor, these clauses are shown in the list on the left and the actual proof (in the form of a tree that is growing from top to bottom) is shown on the right. A proof in Resolution consists of several steps with each step being represented by a line in the proof tree.

A step in Resolution requires two actions: First, clauses from the list on the left have to be selected which are then shown on the current line of the proof tree. Afterward, a literal (one element of a clause) which occurs negated (i.e., a $\neg$ before the letter) in one of the clauses and non-negated (i.e., the same letter but without $\neg$) in the

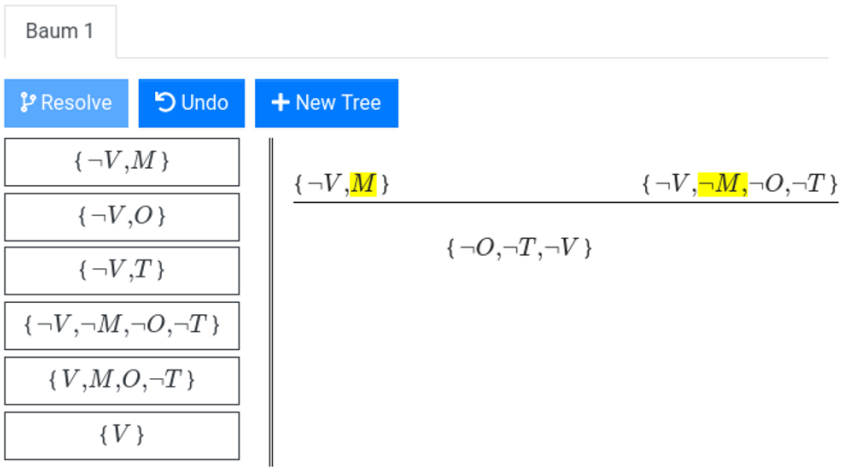


Fig. 4.3.: Problem-specific editor for the proof technique Resolution by example of an exercise on propositional logic (taken from [Sta+19, p. 545]).

other clause has to be selected. After that, a Resolution step can be performed by clicking on the button on the top left labeled *Resolve*. The example in Figure 4.3 shows the proof tree after a Resolution step has been performed: Before that, the two clauses shown above the line were selected, and the literal M in both its negated (right clause) and non-negated (left clause) form has been selected. A Resolution step leads to a new line in the proof tree that already contains a clause which is the union of the two clauses of the previous line minus the selected literal in both its negated and non-negated form. The resulting clause of the Resolution step performed in the example can be seen below the line in which the literal M was selected. After that, the process starts from the beginning with the exception that only one clause has to be selected from the list on the left as there is already one clause present in the current line of the proof tree. That process continues until a Resolution step yields the empty set which shows the unsatisfiability of the set of clauses. The example shows the editor used for a proof in propositional logic but can be used for first-order logic as well in case of which it provides support for variable substitutions and factorization.

To allow users to focus on one of the most important parts of a proof using Resolution, the strategic selection of clauses and literals so that the proof ultimately arrives at the empty set, the editor takes over other parts of the proof, such as actually performing a Resolution step or substituting variables from a user-provided substitution when doing a proof in first-order logic. Throughout working on their proof, users are provided with feedback: Immediate feedback on the correctness of a step is provided through the step being actually performed, that is, a click on the button labeled *Resolve* leads to a new level in the proof tree. Furthermore, feedback on a proof's overall correctness is given immediately after the empty set is reached. Immediate feedback on errors is given through error messages and explanations that are shown

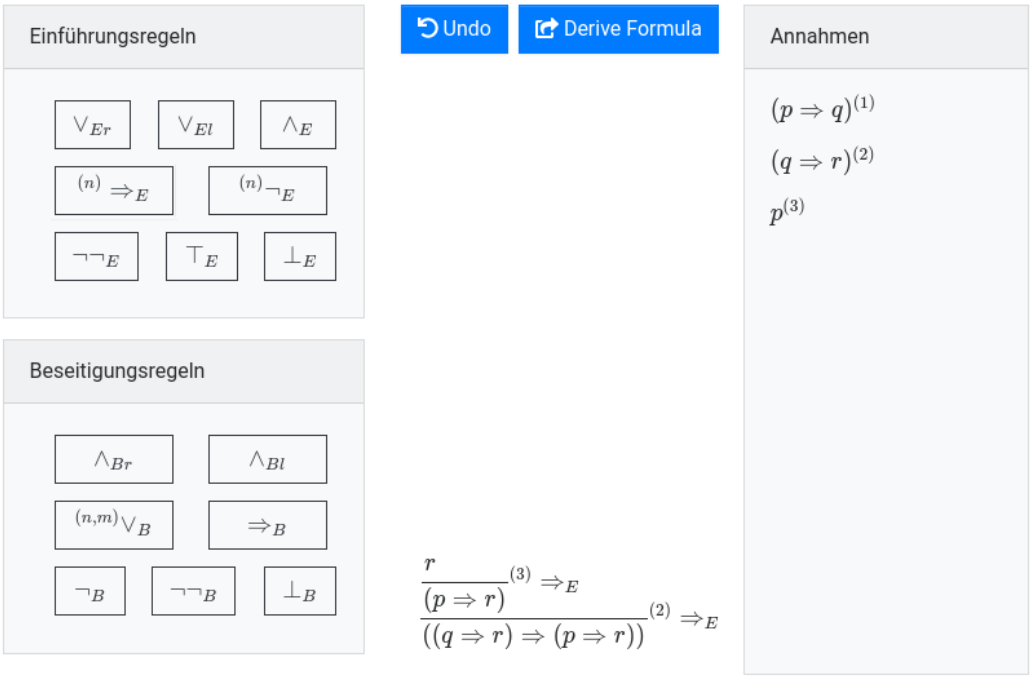


Fig. 4.4.: Problem-specific editor for the proof technique Natural Deduction by example of an exercise on propositional logic (taken from [Sta+19, p. 545]).

after a user tried to perform an impossible step, such as a message stating that a Resolution step could not be performed as exactly one literal in each class has to be selected.

The other proof technique for which an editor was conceived and implemented by Staudacher is Natural Deduction which is used to prove the satisfiability of formulas in propositional or first-order logic [Gen35]. As Natural Deduction is more complex than Resolution, a detailed explanation of the proof technique is forgone at this part; refer to Staudacher's bachelor thesis [Sta18] for such.

The editor for Natural Deduction can be seen in Figure 4.4 and consists of three parts: The left side shows the rules of Natural Deduction, the middle part the current proof tree, and the left side a list of so-called assumptions. Both the list of assumptions and the proof tree change throughout working on the proof, while the rules remain unchanged. A step in a proof using Natural Deduction generally consists of selecting parts of the proof tree and a rule to apply to the selected parts. If the selected rule is applicable to the selected parts, a new level containing the result of the rule application is added to the proof tree. Contrary to the editor for Resolution, the proof tree in the editor for Natural Deduction grows bottom-up, that is, new levels are added on top of previous levels, and can branch. There are, however, situations in which parts of the proof tree can more easily be derived using a top-down approach:

In such situations, these parts can be derived in a separate editor using a top-down approach and afterward be inserted into the proof tree.

Similarly to the editor for Resolution, the editor for Natural Deduction allows users to focus on those parts of the proof technique most important when first being introduced to it: Learning how the rules work, that is, understanding when and to what parts of the proof tree a rule can and should be applied. If the selected rule can be applied to the selected parts of the proof tree, that rule is applied and the proof tree is updated with the result of the rule application. If a rule cannot be applied to the selected parts of the proof tree, an error message and an explanation are shown. Hence, users are provided with immediate feedback on success and errors. Feedback on a proof's overall correctness is provided as well which is especially important for Natural Deduction as various conditions have to be met for a proof to be correct, and hence, it is not always evident for beginners when a proof is finished. Additionally, the editor takes care of assumptions: Assumptions created through rule application are added automatically to the list and removed if they are used by an applied rule.

Implementation in Backstage 2 On the implementation side, Backstage 2 makes a distinction between question types and quizzes: A question type is a class of problems an answer to which can be created with a problem- or subject-specific editor, while a quiz is an instantiation of a question type. Taking choice as an example, the editor is a component that can display a question text and various answer options, while a choice quiz consists of an actual question text and answer options relating to the question text. As already mentioned, quizzes in Backstage 2 are attached to Simple Units, that is, a quiz is always accompanied by a unit.

Three components are required to add a new question type to Backstage 2: An editor which displays the content of a quiz and using which an answer to that quiz can be created, a function that takes the editor's output (i.e., a student's answer) and decides if the answer is a correct answer to a quiz, and optionally a component that can display the model solution to a quiz. How answers are checked for correctness depends on the question type and can be done, for example, by comparing the answer with the correct answer, as in choice, or by examining the answer only, as in Resolution where an answer is correct if the last line of the answer is the empty set.

Figure 4.5 shows the view students are shown during a quiz: In the top part, the unit the quiz is attached to can be seen. Below that, an editor for choice can be seen which shows a question text and several answer options. An arbitrary number of answer options can be selected or unselected by clicking on them. That part, labeled

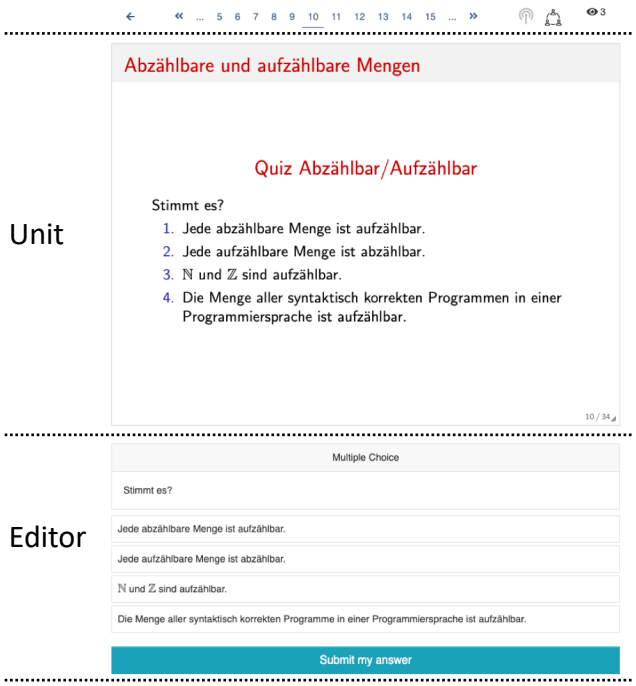


Fig. 4.5.: Student's view while a quiz is running (slide is from François Bry's lecture *Aussagenlogik - Teil 1* licensed under CC BY-NC-SA).

with *Editor* in the figure, is the position where editors are shown regardless of the question type, that is, is the only part that varies between question types.

After an answer has been created and submitted using the button on the bottom right of the figure, the answer is checked for correctness on the server using the aforementioned function. Feedback on correctness (if not already given by the editor) is not given immediately but only after a quiz has been closed by the lecturer, and up until that point, students can change their answers as often as they like.

After a quiz has been closed by the lecturer, every participant receives feedback on the correctness of their answer which can be seen in Figure 4.6: At the top, general correctness feedback is given; in the example, the given answer was correct. If a component for displaying a model solution was provided for the respective question type, the general correctness feedback can be unfolded to show the model solution. The example shows that component for choice which shows the question text, the correct answers, the student's given answers, as well as for each answer option the percentage of students choosing that answer option. However, such a component can be simpler as well and only show the correct answer. The bottom part shows the percentage of correct and incorrect answers given for that quiz.

Even with problem- or subject-specific editors which support students while they work on exercises, not every student might be able to create an answer. Parallel

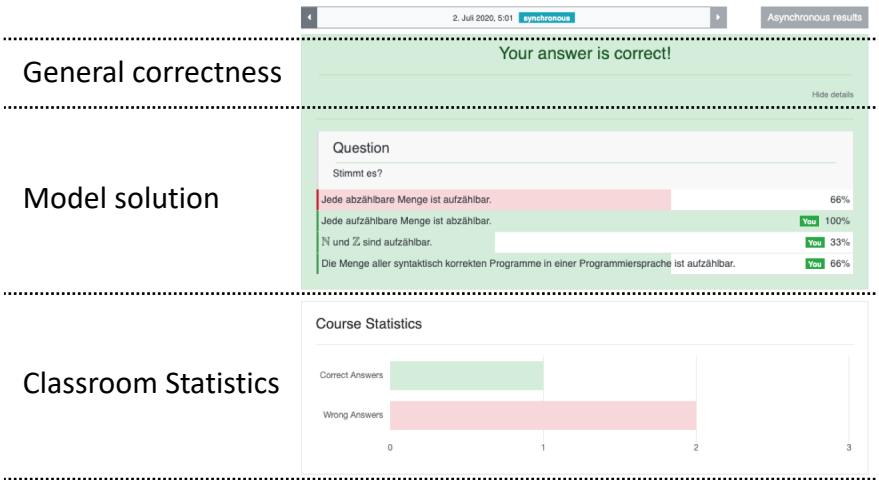


Fig. 4.6.: Student's view after a quiz.

to the support provided to students by the editors, editors and quizzes could be adapted to individual students. How adaptivity could be implemented in audience response systems is shortly outlined in the next section.

4.2.2 Adaptivity

As already mentioned in the previous section, more complex quizzes make it more difficult that every student can produce an answer. The previous section introduced the idea that editors themselves provide support to all students equally while they work on their solution. In this section, an approach running parallel to that support is introduced: Adapting the editor or the quizzes to the individual student so that more students are empowered to produce an answer. The following section shortly outlines ideas for adapting editors and quizzes but does not go into detail as none of the ideas were implemented in Backstage 2's audience response system.

Among the approaches to adapt to an individual user, Hwang [Hwa14] lists to adapt the interface (in this case, the editors) or the materials (in this case, the quizzes). Less experienced students could be shown an easier version of an editor that does more work for them, and quizzes could be adapted through partially-filled editors, that is, showing students the same quiz in different stages of completion. Examples for the latter are the use of the aforementioned editors for logics in which proofs in various stages of completion can be shown, or a coding quiz where only a few lines instead of a complete program have to be written. However, creating partially-filled exercises is associated with high effort, as the lecturer is effectively forced to construct various versions of the same quiz. Determining what parts to show and what parts to leave for users to complete can not always be done automatically, as it depends on the learning goals of a quiz: By the example of a coding quiz, there

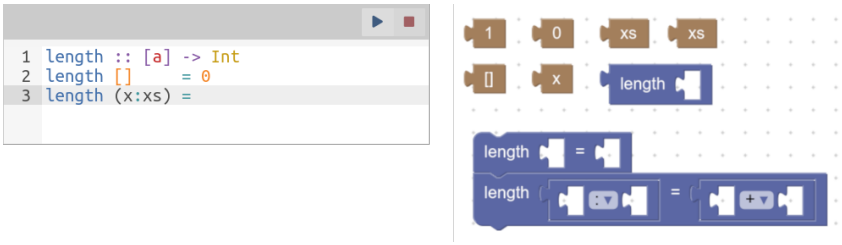


Fig. 4.7.: Two representations of the same quiz: In the left editor, students have to write a whole program on their own, while in they left editor they just connect blocks (taken from [MB19a, p. 211]).

is a difference between letting students just complete variable declarations, the actual program logic, parts of the actual program logic, or just a method header. Furthermore, the same quiz can be used for various learning goals, and for each goal leaving different parts to complete makes sense.

An example of adapting the editor can be seen in Figure 4.7 which shows two editors for the same quiz. In both quizzes, students are tasked to write a program that determines the length of a list in the programming language Haskell. The non-adapted version can be seen on the left side of the figure: Here, students need to write the entire program on their own and are only supported through syntax highlighting, that is, besides having understood the underlying concepts of *pattern matching* and *recursion*, students need to have a grasp on Haskell's syntax as well. In the adapted editor on the right side of the figure, students have to solve the same exercise but require only a rudimentary understanding of Haskell's syntax. The editor was implemented for a very small subset of Haskell using Blockly[4] which allows users to construct programs by connecting blocks using drag-and-drop. A less experienced student might be able to solve the quiz using that editor or at least be able to produce some kind of answer when the same student might have failed to produce anything when supposed to write the entire program from scratch. However, creating various versions of the same editor is associated with even more effort than creating various versions of the same quiz.

Besides the high effort required for the *how* to adapt to a user, another problem has to be solved as well: Deciding *when* to adapt to a user. For that, various data can be considered, such as overall correctness of (similar) quizzes, number of consecutive error messages given by an editor, or time without any action by the user. Adaption can be put into the hands of users as well with users deciding themselves when to change to an easier or more difficult version of a quiz.

[4] `https://developers.google.com/blockly`

As already mentioned, adaptivity never found its way into Backstage 2's audience response system, and hence, this chapter only shortly outlined possible approaches for introducing adaptivity to audience response systems.

4.2.3 Phases

In nearly all of the examined audience response systems, quizzes consisted of two phases: A phase in which students give their answers followed by a phase where the aggregated results are presented to the audience. Adding more phases in between allows lecturers to use an audience response system for more sophisticated classroom interactions, such as letting students review other students' answers before showing the results.

Among the systems examined in the survey, two systems implemented something akin to phases:

- In Informa, a quiz consists of three phases: First, students give their answers, then the aggregated results are shown without indicating what answers are correct, and only in the third phase, the correct answers are revealed [Hau08]. A subsequent version of Informa added another quiz with three phases: To give students who have already finished their answer (to coding tasks) something to do, these students are given answers of other already finished students for review as their second phase [HA13].
- In the audience response system developed by Schön et al. [Sch+16], a quiz can consist of a sequence of quizzes which are done in succession where the lecturer decides when to continue to the next quiz in the sequence.

The understanding of phases as used in this section is insofar different from Informa's second approach that students proceed through phases together controlled by the lecturer, that is, at any time all students are in the same phase. In contrast to the approach by Schön et al. [Sch+16], subsequent phases can be something completely different and not necessarily an instantiation of the same or a different question type. In summary, in Backstage 2's audience response system, a quiz is considered a sequence of an arbitrary number of phases which the participants of a quiz proceed through together controlled by the lecturer. Besides, subsequent phases can use answers or other artifacts generated in the preceding phases.

An illustration of a quiz spanning three phases can be seen in Figure 4.8. In the first phase, students solve a quiz using a problem- or subject-specific editor. As soon as the lecturer decides to proceed to the next phase, each student is assigned

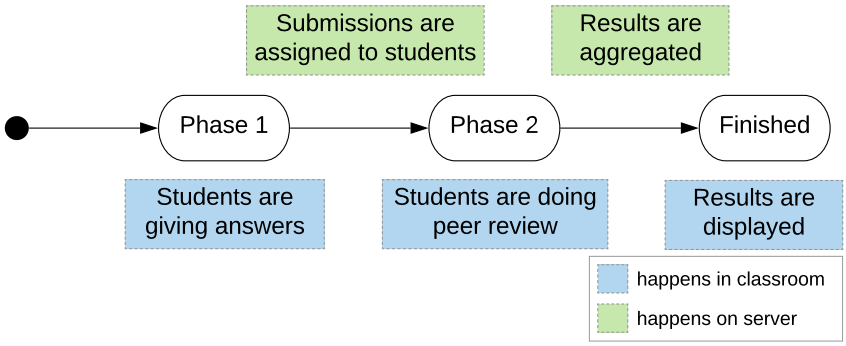


Fig. 4.8.: Quiz spanning three phases: Students first create an answer, then review another student's answer before aggregated results are shown (taken from [MB19a, p. 212]).

another student's submission for review. After that phase, aggregated results are shown as usual. That quiz was implemented for the format Phased Classroom Instruction which used that quiz combined with the JavaScript editor (see Chapter 6). Furthermore, that quiz was implemented to be used with the editors for Resolution and Natural Deduction as well but was only tested in a very small class.

Another quiz with three phases was implemented by Martin Gross [Gro17] as part of his bachelor thesis: Analogously to the previously introduced quiz, students first create an answer using an editor. In the next phase, every user is presented repeatedly with two of their peers' answers and has to decide which one is better (or that both are equal). In the last phase, instead of the usual correctness feedback in the form of aggregated results, a ranking of answers calculated from the students' votes during the second phase is shown. The quiz was intended to be used in programming courses in combination with a coding editor under the assumption that different approaches or misconceptions become visible as the top-ranked answers.

In the teaching method Peer Instruction, students first do a quiz on their own and then discuss their answers with their peers before answering the same quiz again [CM01]. Peer Instruction can be seen as a quiz that consists of three phases: First answer, second answer, and aggregated results. In this case, the aggregated results can show results from both rounds and, for example, visualize how the chosen answers changed between the rounds.

In summary, introducing quizzes that span multiple phases into audience response systems allows to represent more sophisticated classroom instructions that make lecture sessions more engaging and interactive. Phases is the last area outlined as part of this thesis in which audience response systems could grow in the future. The next section shortly outlines various aspects of Backstage 2's audience response system not yet talked about.

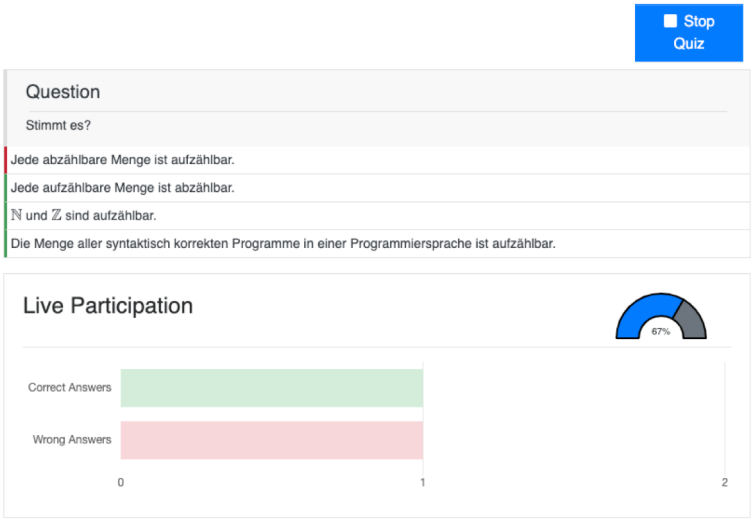


Fig. 4.9.: Lecturers' view while a quiz is running.

4.3 Backstage 2's Audience Response System

An audience response system consists of three parts: What students see, what lecturers see, and what is projected in the lecture hall. What students see during and after quizzes was already introduced as part of Section 4.2.1, and hence, the following section only shortly outlines the latter two parts.

The view lecturers are shown on their own devices can be seen in Figure 4.9: At the top right of the figure, the button for ending the quiz can be seen. Below that, the correct solution to the quiz is shown (depending on whether a component for displaying a model solution exists), and at the bottom, real-time statistics of the percentage of logged-in students already having given an answer and the current percentage of correct and incorrect answers are shown. If intended by lectures, these real-time statistics can be used as a means for intervening while a quiz is still running.

During a quiz, the view shown in Figure 4.10 is projected in the lecture hall: On the top left and right, respectively, the unit and the quiz itself can be seen. The bottom part shows the same real-time percentage of logged-in students having already answered but does not show the current percentage of correct and incorrect answers. Showing the latter could potentially influence students' answers and is therefore omitted from that view.

After a quiz has been closed by the lecturer, the view projected in the classroom changes to the view shown in Figure 4.11. At the same time, students are shown personal feedback on their devices as shown in Figure 4.6. The view projected in the

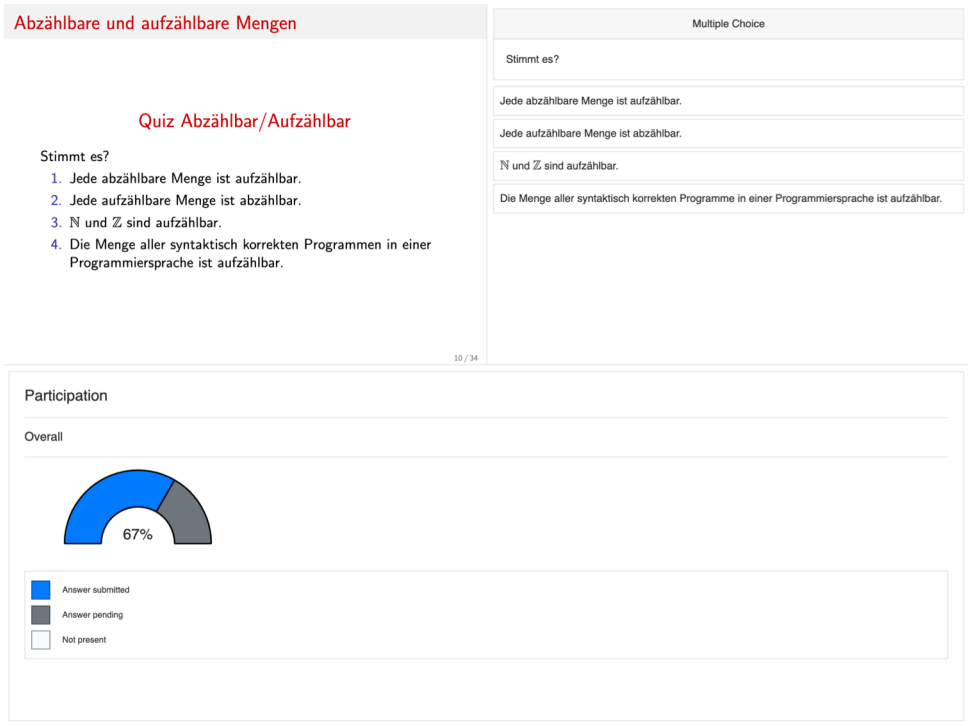


Fig. 4.10.: Projected view while a quiz is running (slide is from François Bry's lecture *Aussagenlogik - Teil 1* licensed under CC BY-NC-SA).

classroom after a quiz provides lecturers a means to discuss the quiz and its results and to that end contains the unit and model solution (if such a component was provided) at the top and the percentage of correct and incorrect answers below that. On their own device, lecturers are presented with the same information differently arranged.

A feature not talked about until now is that Backstage 2's audience response system allows for so-called asynchronous runs of quizzes, that is, students can start and answer quizzes on their own and get immediate feedback. Depending on the configuration of a quiz, the option for asynchronous runs is available never, always, or only after a quiz has been run once through a lecturer during a lecture session. In that way, users can be given the possibility to redo quizzes, for example, when preparing for an examination. Furthermore, support for asynchronous runs makes an audience response system more versatile, as it now can be used for courses with no direct lecturer involvement as well.

4.4 Wrapping up Audience Response System

This chapter introduced Backstage 2's audience response system and the various areas in which it goes beyond what current audience response systems are offering:

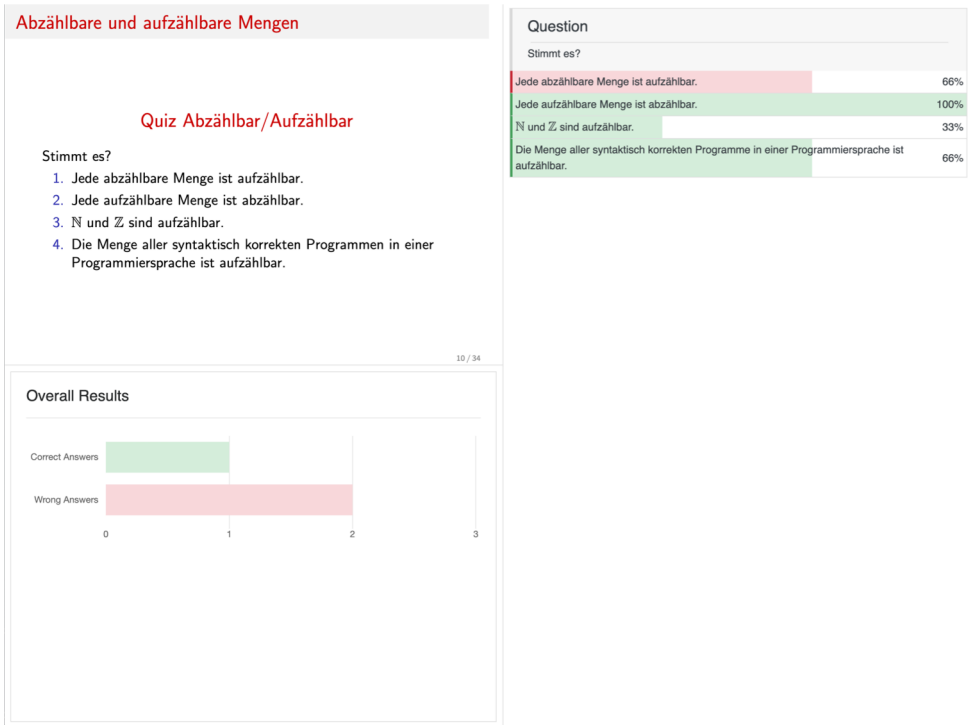


Fig. 4.11.: Projected view after a quiz (slide is from François Bry's lecture *Aussagenlogik - Teil 1* licensed under CC BY-NC-SA).

More complex question types are supported through problem- or subject-specific editors, and quizzes can span more phases than the usual phases which allows for more sophisticated classroom interactions. Furthermore, approaches for adapting quizzes to individual users were discussed which have the goal to empower more students to be able to produce an answer but have not been implemented into Backstage 2's audience response system. Even though current audience response systems are already engaging the audience and introducing interactivity to large courses, both areas in which Backstage 2 extends upon current audience response systems can potentially bring more engagement and interactivity to large lecture halls.

Various editors and a three-phase quiz have been evaluated as part of the learning and teaching formats introduced in Part II. Future work should consider whether adaptivity is conceivable for audience response systems – both regarding the high effort associated with adaptivity as well as learning outcomes of students – and, if so, implement and evaluate various approaches. Furthermore, the editors built for Backstage 2's audience response system mostly focus on STEM education. Future research could consider how such editors – if possible – can look for other subjects.

Most of the learning and teaching formats described in the following part use the audience response system to some extent: In Large Class Teaching, introduced in Chapter 5, the audience response system is used in a conservative way accompanying

a large course with choice quizzes. In Phased Classroom Instruction, introduced in Chapter 6, the previously introduced quiz with peer review spanning three phases was used in combination with the JavaScript editor. In the final format, Bite-sized Learning, described in Chapter 8, a variety of editors are used for two courses with quizzes that are run exclusively asynchronously. Outside of learning and teaching formats, the audience response system was extended with a social gamification based on teams which is introduced in Chapter 9.

Now that the main components – the collaborative annotation system and the audience response system – have been introduced, the next part describes how these, together with the basic elements of Backstage 2, are combined to create four different learning and teaching formats which aim to break the invisible fourth wall between lecturer and audience.

Part II

Breaking the Fourth Wall

The *fourth wall* is "an imaginary wall (...) that keeps performers from recognizing or directly addressing their audience" [Mer20a], and it is broken when performers interact or address their audience.

Expanding this metaphor to lecture halls, there are lecturers as performers, audiences of varying sizes, and an invisible wall between them – sometimes created by the anonymity brought by the sheer size of the audience, sometimes created by pure convention. That invisible wall transcends well beyond the boundaries of the lecture hall with lecturers and students having only few and short opportunities for personal contact.

The following part introduces four technology-enhanced learning and teaching formats, which were created by combining Backstage 2 and its components (see Part I) in various ways, that try to break down this invisible wall to promote engagement and interactivity in tertiary STEM education.

The idea of combining components into learning and teaching formats was first discussed by Niels Heller and the author of this thesis in [Hel+18] and further refined in [Hel+19] where a distinction between core components, teaching methods, and learning and teaching formats was made. A total of six core components were identified through a structured analysis of existing learning management systems. These core components are first combined into teaching methods (e.g., the combination of *Input Interactions* and *Learning Analytics* results in the teaching method *Audience Response*) which are then combined into learning and teaching formats (e.g., *Audience Response* and *Document-based Collaboration* constitute the format *Large Class Teaching*). This thesis deviates from that approach by considering software components with which a respective teaching method can be implemented and combines these software components then into learning and teaching formats: The

collaborative annotation system implements the teaching method *Document-based Collaboration*, and the audience response system, well, implements the teaching method *Audience Response*. In the following, the four technology-enhanced learning and teaching formats built and evaluated as part of this thesis are shortly outlined.

The first format, *Large Class Teaching* (see Chapter 5), combines the collaborative annotation system acting as a backchannel to provide students communication means during lecture sessions and the audience response system to allow lecturers to introduce interactivity into their (possibly large) class.

In *Phased Classroom Instruction* (see Chapter 6), students work on extensive exercises during lecture sessions using problem- or subject-specific editors of the audience response system. At the same time, the lecturer is supported with a real-time overview of students' progress on the exercises to be more easily able to decide who requires their support to be able to solve the exercise.

Collaborative Peer Review (see Chapter 7) uses the collaborative annotation system to provide reviewers and authors access to the same to-be-reviewed document during the review phase. The collective access enables collaboration of various forms between reviewers and authors, such as reviewers addressing disagreements and authors inquiring about reviews.

In the final format, *Bite-sized Learning* (see Chapter 8), students learn on their own with quizzes that leverage the audience response system's various question types. Students are provided with immediate feedback and explanations on the quiz.

Format by format, responsibility for learning shifts from lecturers to students with the lecturer still being present in the background. While it is desirable to give students more responsibility in their learning, that aspect does not invalidate any of the formats that might include more responsibility on the side of the lecturer, as the best possible learning most likely takes then place when the correct format is chosen for the scenario at hand.

5 Large Class Teaching

In this chapter, the collaborative annotation system and the audience response system are combined to promote interactivity and engagement in traditional lectures. That combination was already realized and evaluated with positive results by Alexander Pohl [Poh15] as part of this doctoral thesis. The approach taken here deviates mainly in two aspects from Pohl's: First, the underlying technology, as it is realized using Backstage 2, and second, while Backstage used a backchannel (i.e., restricted its use mainly to lecture sessions), in Backstage 2, a collaborative annotation system is used which can be used equally during and after lecture sessions. This chapter aims to produce further evidence for Pohl's results who found that "[t]he use of Backstage in the four courses can be considered successful" [Poh15, p. 81] and that "Backstage promotes learning-related awareness and activities" [Poh15, p. 81]. Hence, the methods of evaluation are heavily based on those used by Pohl, but new aspects, such as the students' activity on Backstage 2 during and outside of lecture sessions are evaluated as well.

As already discussed in Chapter 1, traditional lectures are often the last resort for teaching in the face of ever-increasing numbers of students and a not corresponding increase in teaching staff. In such large class environments students are passive, and it is hard to introduce interactivity; both aspects which are not conducive for learning. Another aspect discussed in the introduction is that even though the traditional lecture has its drawbacks, it is still an economical and effective way to convey knowledge to a large number of students but is less suited for promoting thought. Hence, Pohl's (and the goal of this thesis), was not to eliminate traditional lectures from higher education but to use technology to address some of the issues traditional lectures face.

During the years, different approaches for addressing the issues of traditional lectures were developed. Carbone [Car99] reports on William Harwood who addresses the issue of passive students and lacking feedback to the lecturer by asking "students to write down any questions they still have concerning the material covered in class. Participation is voluntary, and there are drop boxes stationed around the room" [Car99, p. 40]. Harwood's approach can be seen as a proto-backchannel, an analog version of a technological support often found in today's lecture sessions, which provides "a secondary or background complement to an existing frontchannel"

[Yar06, p. 852]. Backchannels provide students means for asking questions and making remarks during lecture sessions (the front channel) which then can either be answered by lecturers or their peers [Poh15]. As already discussed in Chapter 4, audience response systems are technology that allows lecturers to conduct (and get the results of) classroom quizzes even with a large number of students.

This chapter first introduces the previous version of Backstage and summarizes the results of its evaluation, and then introduces and discusses the results of two courses where Backstage 2 was used with the format Large Class Teaching. Finally, this chapter lists several improvements to Backstage 2 which could improve the learning format.

5.1 Backstage Then

The first version of Backstage was an educational software which consisted of a backchannel and an audience response system. The following section is a summary of the description and the evaluation of Backstage found in [Poh15]. Note that a few aspects mentioned here were already discussed in Part I, but are repeated here for the reader's convenience.

After joining a lecture session on Backstage, students are presented with a view similar to the screenshot shown in Figure 5.1. In the middle, the current slide is shown. The bar at the top provides access to various functionalities, such as the option to synchronize with the lecturer which leads to a student's current slide changing automatically when the lecturer changes their current slide.

Figure 5.1 shows three backchannel posts as well: Two are referring to positions on the current slide and are shown additionally to their textual representation on the left by icons on the slide. By anchoring backchannel posts to parts of the slide, posts can easily be given a context. The third backchannel post is an answer to an existing post, which is indicated by the text *A reply to post by ...* before the content of the post.

As already discussed in Chapter 3, for creating backchannel posts, Pohl devised a three-step process: Users first click on the position on the slide they want their post anchored to, whereupon they have to select a purpose for their post, before being able to enter the content.

Clicking on a backchannel post on the left reveals further information and interaction possibilities: Users are shown the current rating of the post, can rate and create a

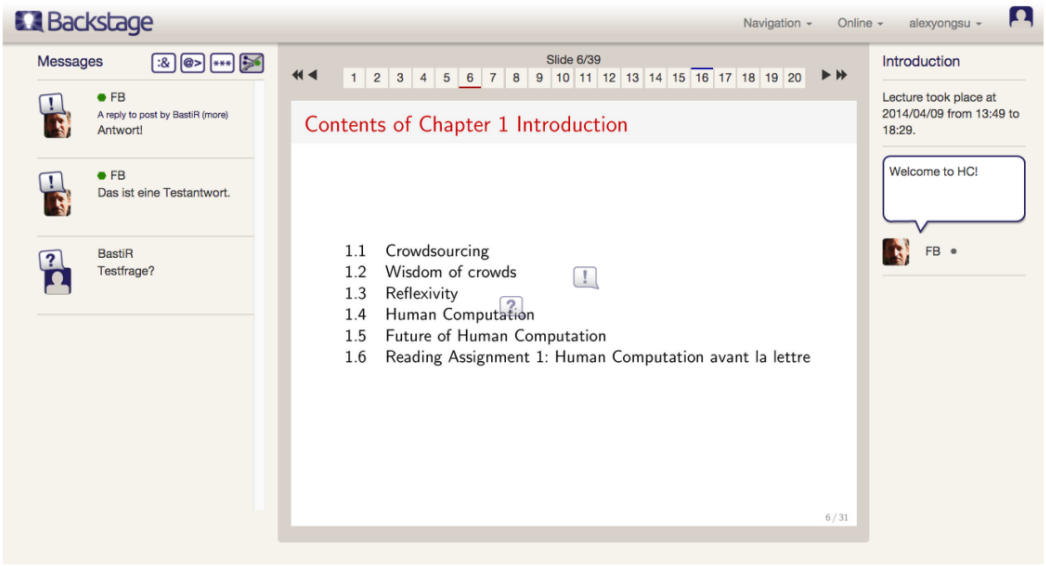


Fig. 5.1.: Active lecture session in the first version of Backstage: In the middle, the lecture slides are shown. On the left, backchannel posts referring to positions on that lecture slide are shown (taken from [Poh15, p. 44]).

reply to the post. Replying to a post is the second possibility to create a backchannel post. The anchor of a reply is the backchannel post the user is replying to. Through rating, users can agree and disagree with posts, as well as mark posts as off-topic.

At any point in a lecture session, lecturers can start a quiz which is shown to students with a view similar to the screenshot shown in Figure 5.2. On the right, the question to be answered and the possible answer options are displayed. On the left, the user can choose one or more of the presented answer options. After a quiz has been run, overall classroom results are displayed. In addition to multiple choice quizzes, the first version of Backstage had support for polls (i.e., quizzes with no incorrect answer) and questions that could be answered with free text.

The first version of Backstage was evaluated in four courses with a large number of students. The evaluation was done using data taken directly from the Backstage system, such as login data, backchannel posts, and participation in classroom quizzes, as well as a survey conducted during the last lecture session of each course. The backchannel posts were classified using a coding scheme adapted from [Cog+01]. The possible categories for a post were *content-oriented*, *organizational*, *process-oriented*, *participation-enabling*, and *independent*. The survey contained several Likert items using which the constructs "INTERACTIVITY, RATING, AWARENESS, and REWORK" [Poh15, p. 67] were measured.

From the surveys' results, as well as the active backchannel communication and quiz participation, Pohl concluded that "[t]he use of Backstage in the four courses can

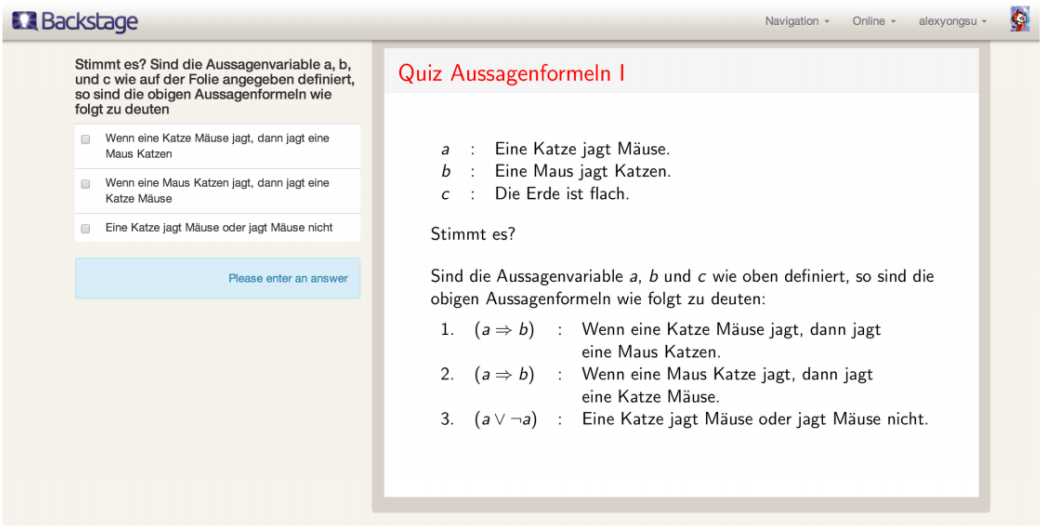


Fig. 5.2.: Student's view of a running quiz in the first version of Backstage: On the right, the question to be answered in shown, on the left, a student can select on or more answer options (taken from [Poh15, p. 52]).

be considered successful" [Poh15, p. 81]. Pohl also reports on the use of Backstage for rework indicated by positive results to REWORK and a large number of logins to the system outside of lecture sessions. Additionally, Pohl concludes that "Backstage promotes learning-related awareness and activities" [Poh15, p. 82], as the surveys show positive values towards AWARENESS and the backchannel communication being mainly content-oriented. Refer to Pohl's thesis [Poh15] for a more detailed view on the results of the evaluation of the first version of Backstage.

The collaborative annotation system, which shares many similarities to the backchannel of the first version of Backstage, and the audience response system of Backstage 2 can be used to provide students a similar environment as the first version of Backstage and constitute the format *Large Class Teaching*. The following section introduces and discusses the results of evaluations of Backstage 2 in two courses.

5.2 Study

The use of Backstage 2 using the format Large Class Teaching was evaluated in two venues of the same course on logics and discrete mathematics. The following section first introduces the courses, then talks about the methods used for evaluation, and concludes with the results and their discussion.

5.2.1 The Courses

Backstage 2 was evaluated in a course on logics and discrete mathematics which is split into two parts: The first six lecture sessions focus on logics and cover among other propositional logic, first-order logic, and the proof techniques Resolution and Natural Deduction. The following four lecture sessions are on discrete mathematics and cover natural numbers, whole numbers and primes, modular arithmetic, and combinatorics. Each part is concluded by a lecture session dedicated to an exemplary examination covering the topics discussed in the preceding part. Therefore, the first part spans a total of seven lecture sessions and the second part a total of five lecture sessions, with the whole course spanning twelve lecture sessions.

Due to technical issues, Backstage 2 could not be used for the entire duration of the course in both venues: In the first venue, **LC1**, the lecture hall initially assigned for the course was too small for the audience and had insufficient WiFi coverage which prevented students and lecturer from reliably accessing Backstage 2. As a better-equipped lecture hall was available later in the term, Backstage 2 was officially reintroduced starting with the seventh lecture session, the exemplary examination for the part on logics. In the second venue, **LC2**, a programming error prevented Backstage 2 to be reliably accessed during the first three lecture sessions. After the error was fixed, Backstage 2 was available reliably starting from the fourth lecture session.

5.2.2 Methods

For evaluation, data taken directly from Backstage 2 and data gathered using a survey was used.

Survey In **LC1**, the survey was conducted online after the examination at the end of the course where only a few students participated. To improve the response rate in **LC2**, both a paper survey during the last lecture session as well as an online survey were conducted. Not all parts of the survey were evaluated as part of the evaluations below. The following lists only those parts which were used for the evaluations. The entire survey can be found in Appendix A.1.

1. 34 items to be rated on a Likert scale to measure the constructs INTERACTIVITY, RATING, REWORK, and AWARENESS adapted from Pohl's [Poh15, p. 163–169] evaluation of the first version of Backstage.
2. The System Usability Scale (SUS) (see [Bro+96]) for measuring the usability of Backstage 2.

3. 5 questions to be answered with free text asking for what students liked / disliked most about Backstage 2, possible improvements for Backstage 2, and opinions towards the engaging aspects of Backstage 2.

All data measured using Likert scales used a scale from *strongly agree* (which was assigned the value 6), to *strongly disagree* (which was assigned the value 1) with no neutral choice.

For each of the constructs INTERACTIVITY, RATING, REWORK, and AWARENESS three to six of the Likert items of (1) were taken and averaged for each participant. From that list, the median was calculated which represents the score for the respective construct. The assignment of items to constructs was adapted from Pohl [Poh15]; refer to Appendix A.1 for a detailed listing of which items constitute which construct.

Data Extraction from Backstage 2 To assess the activity on Backstage 2 during the term, different events were extracted from the system's database. Students could interact in various ways with Backstage 2 with the majority of those interactions resulting in an artifact with a timestamp (hereafter called *activity event*) saved in the database:

- Each time a user *navigated to a Compound Unit*, an artifact was created in the database.
- Each time a user *navigated to a Simple Unit within a Compound Unit*, an artifact was created in the database. Thus, for a user who completely browsed a Compound Unit consisting of 21 Simple Units, 21 artifacts would have been created in the database.
- Each *quiz response* was saved in the database the moment the user pressed the submit button. The timestamp of a response, therefore, represents the moment the user finished working on that quiz. A response was saved regardless of the quiz being conducted during a lecture session or worked on by the student on their own. Overwriting an already given answer in a classroom quiz led to the creation of a further response.
- Each *annotation* was saved in the database the moment the user created that annotation.
- Each *read of an annotation* was saved in the database. An annotation was seen as read either when the user clicked on that annotation or hovered more than two seconds with the mouse over the annotation. Thus, the number of reported reads underestimates the reads that actually took place, as users could read annotations without fulfilling any of those two conditions.

- Each *rating of an annotation* was saved in the database the moment the user rated that annotation.

All activity events for each course were collected and afterward for each activity event determined whether it took place during a lecture session or outside a lecture session. For determining whether an activity event took place during a lecture session, it was checked whether it occurred between begin and end of the lecture session rounded to the whole hour. In other words, if a lecture session was scheduled to be held on a certain day between 11 a.m. and 2 p.m. (which, in Germany, would mean that the lecture session would run from 11:15 a.m. to 1:45 p.m.), an event was seen happened during the lecture session if it occurred on that day between 11 a.m. and 2 p.m. After that, all events were binned by day and for each bin, the number of users doing at least one action (i.e., for which at least one activity event exists) was determined.

Significance was determined using the Mann-Whitney U test, as the majority of data does not follow a normal distribution which calls for a non-parametric test (see [CF14]). The significance threshold was set to $p = 0.05$. Aggregated measures are reported as Median, hereafter abbreviated as Mdn, as it is more robust against outliers [How09], and therefore, deviation is reported as Median Absolute Deviation, hereafter abbreviated as MAD (see [RC93]).

5.2.3 Results

In the following, the results of both courses are introduced. First, general information about the population of the courses is given, and then general activity on Backstage 2 throughout the term is shown. Afterward, activity is broken down into two areas: First, usage of the collaborative annotation system during and outside of lecture sessions, and second, the usage of the audience response system both for participating in classroom quizzes as well as for answering quizzes outside the lecture sessions.

Population of the Courses Table 5.1 shows the number of students and the number of survey participants for both courses. The number of students is nearly identical across both venues, but in **LC2** the number of survey participants was over three times greater than the number of survey participants in **LC1**. Despite that, the response rate of the survey in **LC2** still represents only a small fraction of the whole audience of the course.

Tab. 5.1.: Overview of the population of the courses in which Large Class Teaching with Backstage 2 was evaluated.

Course	Year	# of participants	# of survey participants
LC1	2018	614	17
LC2	2019	609	55

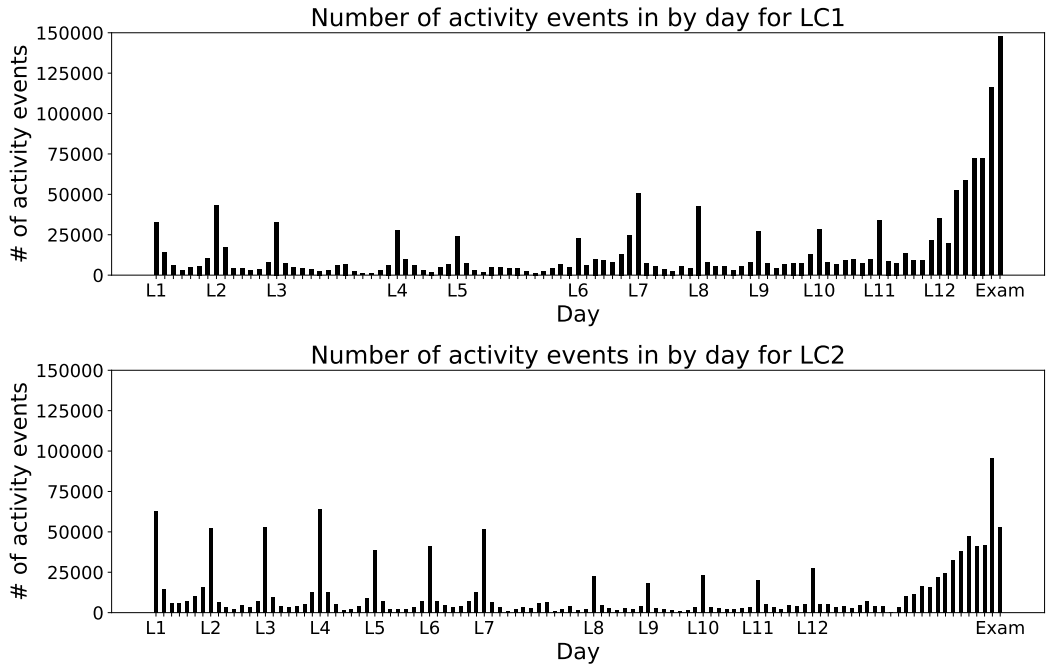


Fig. 5.3.: Number of activity events by day for **LC1** and **LC2**. Each bar represents one day. Labels on the y-axis represent the respective lecture session and the examination.

Activity on Backstage 2 Figure 5.3 shows all activity on Backstage 2 throughout the term, with each bar representing one day. Note that this, and all following figures, show the whole term from the first lecture session to the examination. The labels on the y-axis represent the days on which lecture sessions or the examination took place, respectively. In both courses, days on which lecture sessions took part constituted the peek of that week, with activity beginning to rise the days before the lecture session and dropping the days after the lecture session before beginning to rise again approaching the next lecture session. Generally speaking, across both venues, the activity for each lecture week presents itself as a small normal distribution with the day of the lecture session being the peak. Towards the examination, activity on the system began to increase in both courses. Activity events per day in **LC2** (Mdn: 4677, MAD: 3772) were generally lower than activity events per day recorded for **LC1** (Mdn: 6956, MAD: 4267).

The same pattern as for the absolute number of activity events can be seen when looking at the number of unique active users by day shown in Figure 5.4. Generally, the number of unique users by day was lower in **LC2** (Mdn: 45, MAD: 37) compared

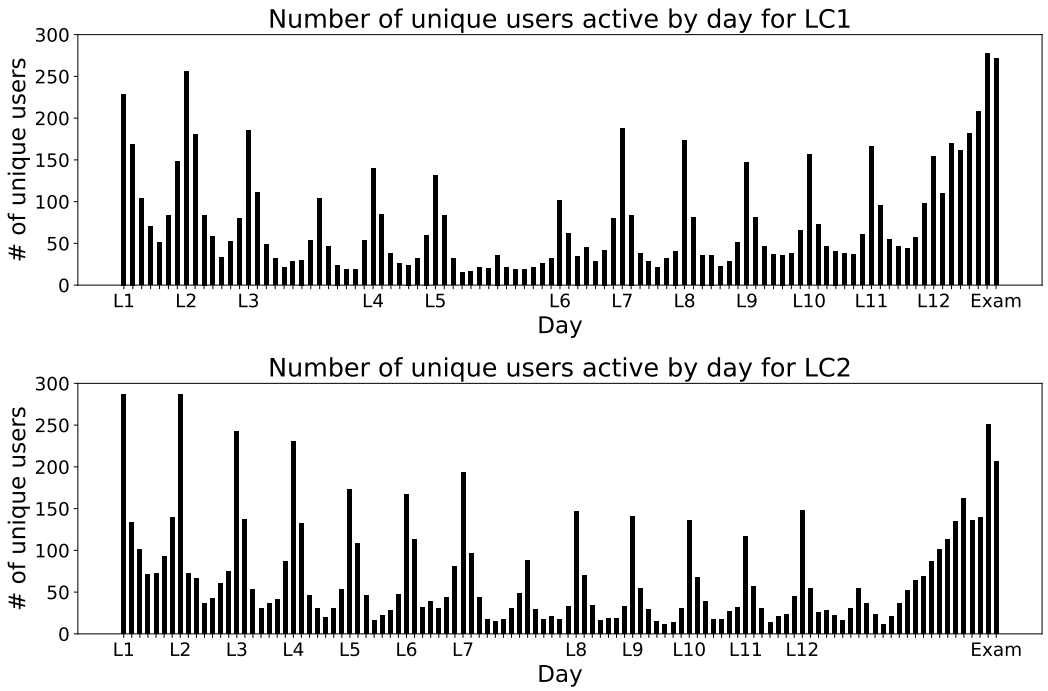


Fig. 5.4.: Number of unique users interacting at least once with Backstage 2 by day for **LC1** and **LC2**. Each bar represents one day and labels on the y-axis represent the respective lecture session and the examination.

to **LC1** (Mdn: 51, MAD: 40). Another observation that can be made from the figure is that even on weeks where no lecture session was scheduled (between L3 and L4 and between L5 and L6 in **LC1** and between L7 and L8 in **LC2**), there was still a peak on the day on which the lecture session would have taken place, with the remaining days of the week following the aforementioned pattern of increasing before and decreasing after the lecture session.

The results for the absolute number of events and unique users by day suggest that the usage of Backstage 2 followed a similar pattern each week. Patterns in time series can be identified using the autocorrelation $\rho(k)$, that is, correlating time series data with the same time series shifted by k (also called lag): Plotting the autocorrelation for increasing values of k yields diagrams in which patterns in the time series become evident [Bro06]. Figure 5.5 shows such plots for the unique users by day. k was increased consecutively by one (i.e., representing a shift by one day). Labels on the horizontal axis represent those values for k on which the shifted time series was shifted to the day of the labeled lecture session. Thus, the label L2 represents the original time series correlated with a version shifted by seven days, the day the second lecture session took part.

The graphs that can be seen in Figure 5.5 confirm that the unique users by day followed the same pattern each week, with being significant (assuming the 99% confidence threshold) the first three weeks in **LC1** and the first five weeks in **LC2**.

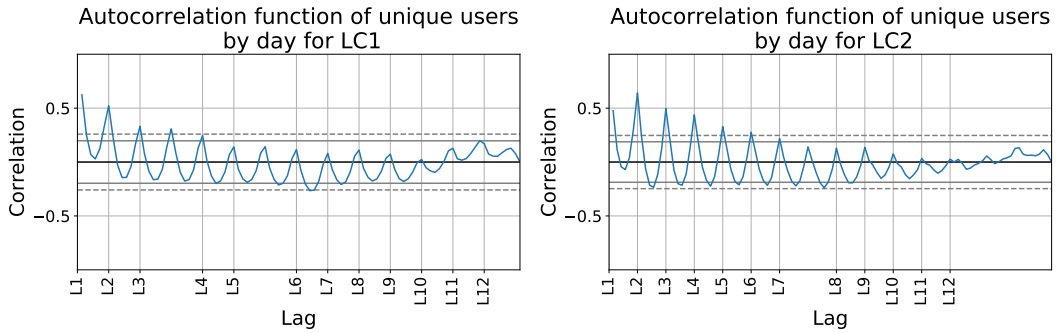


Fig. 5.5.: Autocorrelation function for of unique users by day for **LC1** and **LC2**. Lag was consecutively increased by one. The dotted line represents the 95% confidence interval; the straight line the 99% confidence interval.

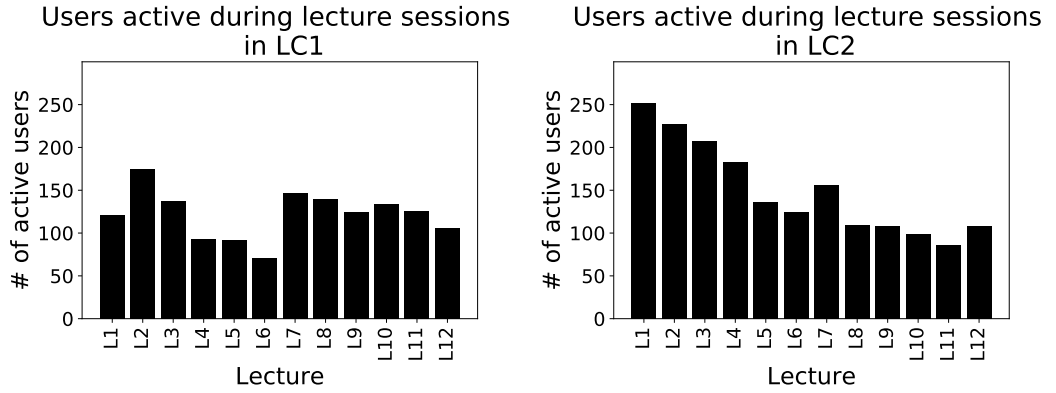


Fig. 5.6.: Number of users active by lecture session for **LC1** and **LC2**.

While no longer significant, the pattern still was clearly visible in the following weeks, before no longer following the pattern at the end of the term going towards the examination.

The number of users active during lecture sessions can be seen in Figure 5.6. Although Backstage 2 was not used during the first part of the course in **LC1**, there were users active on Backstage 2 during the lecture sessions (Mdn: 121, MAD: 42). In the second part of the course in which Backstage 2 was used for classroom quizzes, the number of users active during lecture sessions was slightly higher and more consistent (Mdn: 126, MAD: 3), but slowly dropped towards the end of the term. With two outliers, the number of users active during lecture sessions dropped in **LC2** throughout the term. In **LC2** there were more users active during lecture sessions in the first part (Mdn: 183, MAD: 65) than in the second part of the course (Mdn: 99, MAD: 15). For both courses, the number of active users during lecture sessions was significantly different between the two parts of the course ($p = 0.015$ for **LC1**, $p = 0.006$ for **LC2**). Furthermore, the difference in the number of active users during lecture sessions in the second part of **LC1** was significantly higher than during lecture sessions in the second part of **LC2** ($p = 0.003$).

Tab. 5.2.: Aggregated numbers of unique users of backchannel functionalities during lecture sessions for **LC1**.

	First part		Second part		Overall	
	Mdn	MAD	Mdn	MAD	Mdn	MAD
# of annotators	5.0	3.0	4.0	1.5	4.5	2.2
# of voters	9.0	5.9	5.0	5.9	7.0	5.9
# of readers	34.0	14.8	18.0	13.3	25.5	11.9

Tab. 5.3.: Aggregated numbers of unique users of backchannel functionalities during lecture sessions for **LC2**.

	First part		Second part		Overall	
	Mdn	MAD	Mdn	MAD	Mdn	MAD
# of annotators	12.0	4.4	3.0	1.5	6.0	5.9
# of voters	9.0	10.4	2.0	0.0	4.5	5.2
# of readers	45.0	37.1	8.0	1.5	14.5	11.9

In the following sections, the usage of the collaborative annotation system and the audience response system is examined in greater detail.

Usage of the Collaborative Annotation System The collaborative annotation system allowed users to create either public or private annotations. Both forms of annotation could take place either during lecture sessions or outside of lecture sessions. In the following, the occurrence of each of those usage patterns and, in case of public annotations, how other users interacted with those annotations, is discussed.

The use of the collaborative annotation system during lecture sessions can be seen as using the collaborative annotation system as a backchannel. Tables 5.2 and 5.3 show an overview of the number of unique users using the respective feature during lecture sessions for **LC1** and **LC2**, respectively. A user was seen as using a feature if they performed at least one action of that type during a lecture session. Across all features and for both courses, more students were active in the first part of each course than the second part of that course. Compared to the previously reported number of active users during lecture sessions, only a small proportion of users were engaging with the backchannel. Users who read backchannel posts were the largest group of those interacting with the backchannel but were still only a fraction of the total number of users being active on the system during lecture sessions.

The same numbers broken down into individual lecture sessions can be seen in Figure 5.4 for both courses. Additionally, the number of users active during lecture sessions is displayed as a grey line to allow for easier comparison. The observation that only a small proportion of users actively participated in backchannel communication is

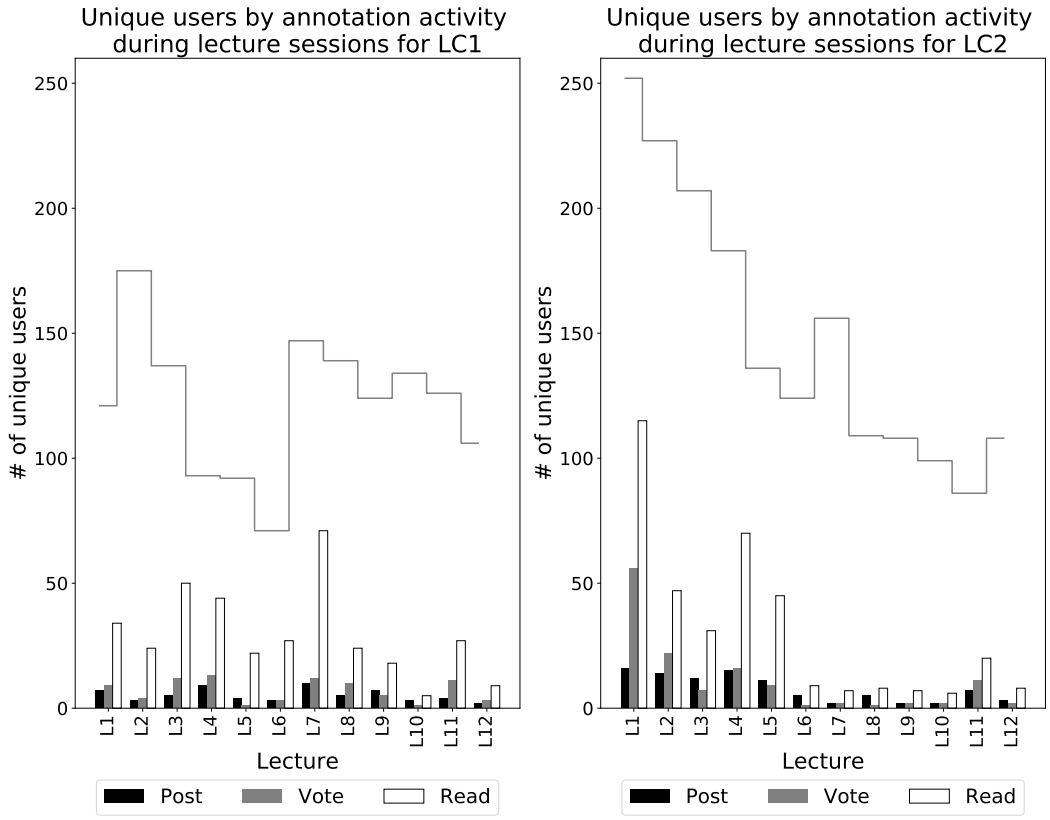


Fig. 5.7.: Overview of unique users using a certain feature of the collaborative annotation system during lecture sessions for **LC1** and **LC2**. The number of users active during each lecture session is indicated by the grey line.

evident in that figure as well: With exceptions, even the number of users who read at least one annotation never reached half of the users active during that lecture session. Furthermore, the stark contrast between the first and second parts of **LC2** can be seen in Figure 5.4 as well. In contrast, the backchannel participation in **LC1** was more consistent throughout the term.

The collaborative annotation system was used outside of lecture sessions as well: Table 5.4 shows an overview of the number of students using the collaborative annotation system outside of lecture sessions grouped by feature and Figure 5.8 shows the same data broken down into individual lecture weeks. In both courses, more students engaged with the collaborative annotation system outside of lecture sessions than during lecture sessions. Interaction with the collaborative annotation system was generally more consistent throughout the term in **LC1** than in **LC2**. In **LC2** the drop in activity between the two parts of the course becomes evident in the usage outside the lecture sessions as well.

Lastly, some students used the collaborative annotation for creating private annotations. Table 5.5 shows an overview of the number of users creating at least one private annotation. In both courses, students used the collaborative annotation

Tab. 5.4.: Aggregated numbers of unique users of collaborative annotation system functionalities outside of lecture sessions for **LC1** and **LC2**.

	LC1		LC2	
	Mdn	MAD	Mdn	MAD
# of annotators	11.0	5.9	5.0	4.4
# of voters	17.0	4.4	17.0	14.1
# of readers	79.0	20.8	56.0	39.3

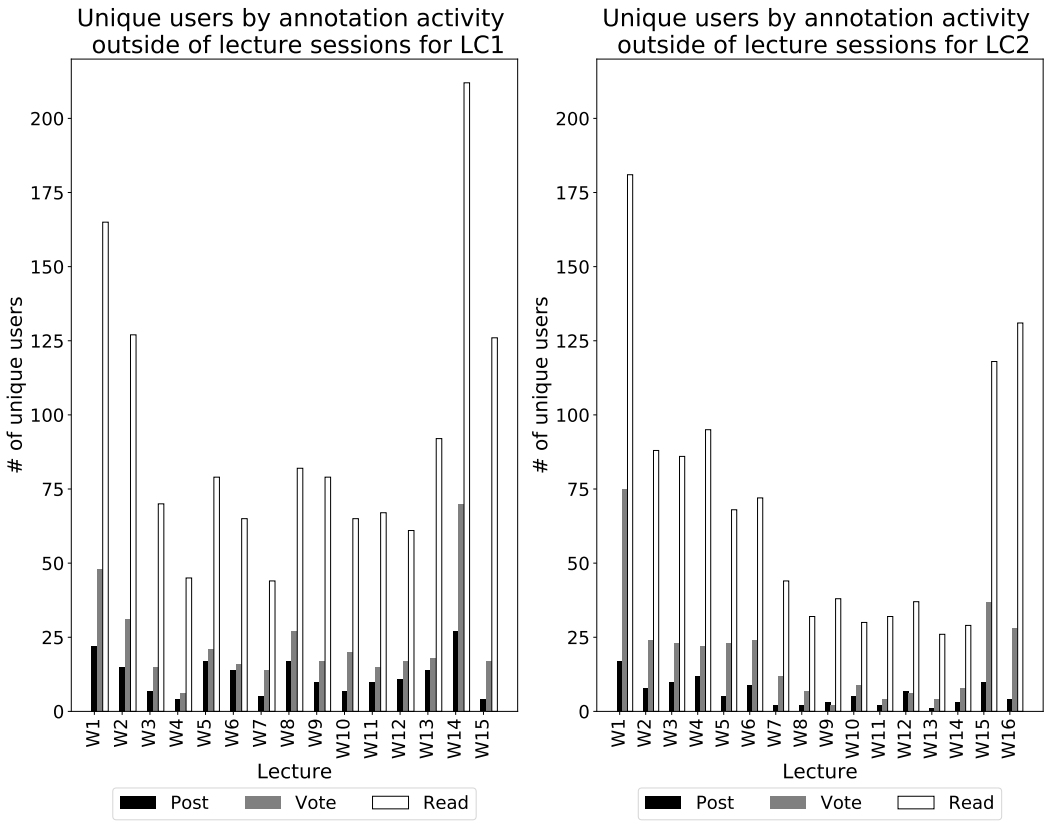


Fig. 5.8.: Overview of unique users using a certain feature of the collaborative annotation system outside lecture sessions by week for **LC1** and **LC2**.

Tab. 5.5.: Number of users creating private annotations during and outside of lecture sessions for **LC1** and **LC2**.

		# of annotators			
		Mdn	MAD	Max	Min
LC1	During class	4.0	1.5	5	0
	Outside class	11.0	5.9	27	4
LC2	During class	4.0	3.0	14	2
	Outside class	5.0	4.4	17	1

system to create private annotations, both during as well as outside of lecture sessions. The numbers of users creating private annotations during lecture sessions were similar in **LC1** and **LC2**, with the number of private annotators staying more consistent in **LC1** than in **LC2**.

The audience response system was the second component that was used to promote interactivity and engagement in large classes and its use by students over the course of each term will be discussed in the next section.

Use of the Audience Response System The audience response system could be used in two ways: During lecture sessions in the form of classroom quizzes and outside of lecture sessions by students for doing quizzes on their own accord.

Figure 5.9 shows the number of users participating in at least one classroom quiz during the respective lecture session and the total number of users active for each lecture session. In the reported results for **LC1** only the lecture sessions starting from the seventh lecture session were taken into account, as that was the first lecture session in which Backstage 2 was used for classroom quizzes.

Overall, in lecture sessions where Backstage 2 was used for classroom quizzes, students participated in similar magnitudes in **LC1** (Mdn: 98.5, MAD: 6.7) and **LC2** (Mdn: 96.0, MAD: 44.5) with students in **LC1** participating more consistently compared to the students in **LC2**. With few exceptions, the number of participating students steadily decreased throughout the term in **LC2**. In **LC1** the number of students increased each lecture session after the introduction of classroom quizzes and only dropped for the last lecture session. The number of students participating in classroom quizzes in the second part of the course in **LC2** (Mdn: 67, MAD: 5.9) was lower and less consistent than in the second part in **LC1** (Mdn: 101, MAD: 4.4).

The asynchronous use of the audience response system in both courses can be seen in Figure 5.10. In **LC1**, asynchronous quizzes were only available starting from the

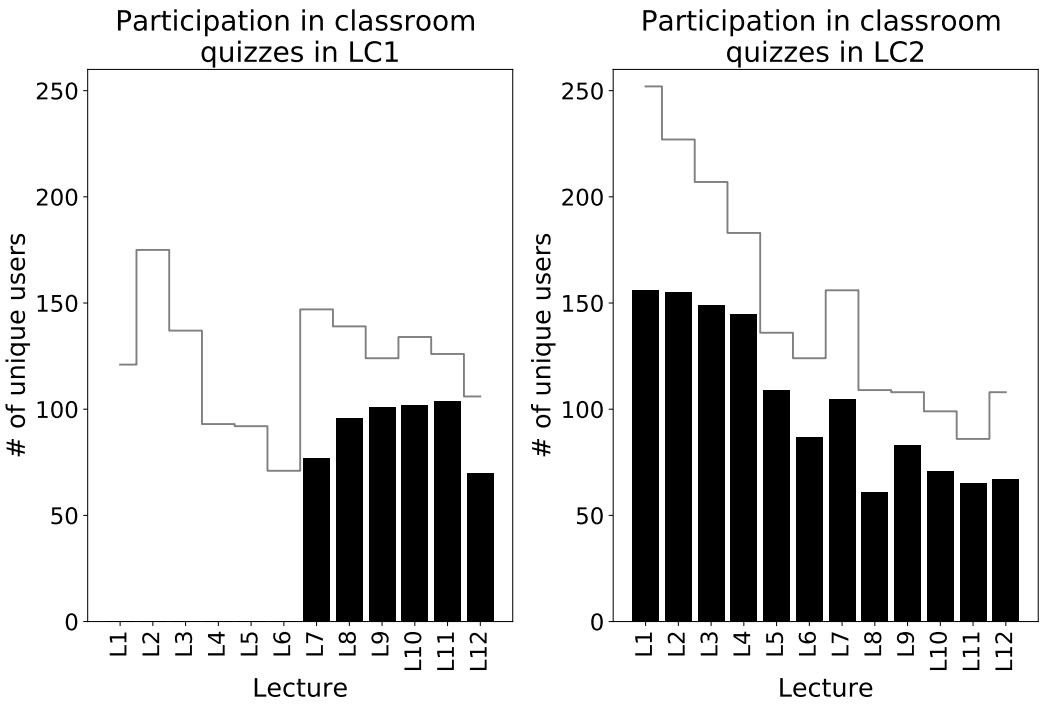


Fig. 5.9.: Number of users participating in at least one classroom quiz by lecture for **LC1** and **LC2**. The grey line indicates the total number of active users during the respective lecture session.

last lecture session, and therefore, there is no activity the weeks before. In **LC2**, the number of users doing asynchronous quizzes showed the previously observed pattern as well: Quiz activity begins to pick up towards the lecture session, peaks on the day of the lecture session, and drops from there on gradually before beginning to pick up again towards the next lecture session. That pattern is visible in the week where no lecture session took part as well. In both courses, the number of users doing asynchronous quizzes strongly increased after the last lecture session towards the examination but was generally lower in **LC2**.

After discussing data gathered from the system in the previous three sections, the final section will discuss data gathered from the surveys.

Students' Attitude towards Backstage 2 For assessing students' attitudes towards Backstage and its effects, Pohl [Poh15] measured the four constructs INTERACTIVITY, RATING, REWORK, and AWARENESS. Each construct was assigned several Likert items from the survey, which, taken together, result in a score for the respective construct. Additionally, the surveys used in **LC1** and **LC2** measured the usability using the System Usability Scale. The following section introduces the results for the four constructs for **LC1** and **LC2**, compares the results to Pohl's results, and finally discusses the results of the usability survey. The surveys used in Pohl's evaluations and the surveys used in the evaluations described here were not completely identical,

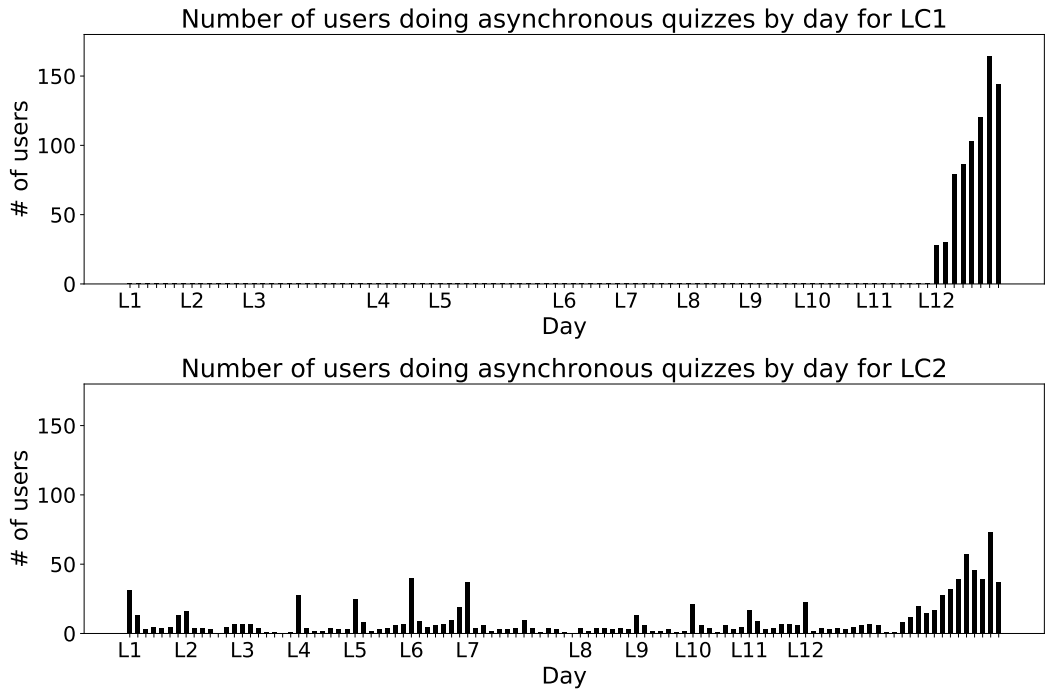


Fig. 5.10.: Number of users doing at least one asynchronous quiz by day for **LC1** and **LC2**.

Tab. 5.6.: Overview of the constructs measured by the surveys (caption and descriptions taken verbatim from [Poh15, p. 68], α replaced with values for **LC1** and **LC2**.)

Construct	Description	α
INTERACTIVITY	measures the usefulness of Backstage as a means to promote interactivity in lectures	0.72
RATING	measures the students' assessments of rating to mark relevant backchannel comments	0.85
REWORK	measures the usefulness of Backstage as a means to gather learning related awareness	0.80
AWARENESS	measures the usefulness of Backstage for reworking lectures	0.65

and therefore, the items which measured a construct had to be adapted. A detailed overview of the items used to measure a concept in [Poh15] and the items used to measure the construct in the evaluations described here can be found in Appendix A.1.

Table 5.6 shows an overview of the constructs, their description, and the value for Cronbach's α for each of the constructs treating the responses to both courses as one. Cronbach's α is a measure for the internal consistency of a scale where a value above 0.7 can be considered as satisfactory [BA97]. In his evaluation, Pohl reported on values for α of above 0.82 for all constructs. Therefore, scale consistency was generally lower for the evaluations described here but except for AWARENESS still satisfactory.

Tab. 5.7.: Measured values for each of the constructs for LC1, LC2, and for both courses.

	P (n = 38)		L (n = 18)		P1 (n = 51)	
	Mdn	MAD	Mdn	MAD	Mdn	MAD
INTERACTIVITY	5.25	0.62	5.42	0.62	5.33	0.49
RATING	4.33	0.99	3.83	1.24	4.67	0.99
REWORK	4.75	1.11	5.12	0.93	3.75	1.11
AWARENESS	5.00	0.59	5.30	1.04	4.80	0.89

Tab. 5.8.: Measured values for each of the constructs for LC1, LC2, and for both courses.

	LC1 (n = 15-17)		LC2 (n = 49-50)		Overall	
	Mdn	MAD	Mdn	MAD	Mdn	MAD
INTERACTIVITY	5.0	0.49	4.67	0.49	4.83	0.74
RATING	4.0	0.99	4.33	0.99	4.33	0.99
REWORK	4.75	0.74	4.75	1.11	4.75	0.80
AWARENESS	5.0	0.59	4.6	1.19	4.6	0.89

To provide a baseline, Table 5.7 gives an overview of the values measured by Pohl for the constructs. The first version of Backstage was generally rated positively across all constructs, with INTERACTIVITY and AWARENESS being rated highest across all three courses. Use of Backstage for REWORK was rated high in **P** and **L** and less, but still positive, in **P1**. Across all courses, RATING was the lowest-rated construct, but even in the course where the construct was rated lowest, **L**, it was still rated positively.

The values for the constructs measured in **LC1** and **LC2** can be seen in Table 5.8. The general trend across both courses seems to be that Backstage 2 was rated a bit more negative than the first version of Backstage, but that all values are still being on the positive side.

The surveys in **LC1** and **LC2** included questions to be answered with free text as well. In the following, trends in the responses given by the students in **LC1** and **LC2** are reported. Note that the following summary of students' answers to those questions is not a formal content analysis, but an identification of trends done by the author of this thesis. The first question asked students *What are for you positive aspects of Backstage?*. Across both courses, 46 students provided a response. Of those 46 students, 20 mentioned features provided by the audience response system as notable positive features, which was expressed by statements similar to the following:

- "Quizzes were awesome and kept me engaged"
- "the quizzes really motivate to be attentive during the lecture"

- “The quizzes really helped me to identify problems i had, so I could focus on them.”

14 of the responses referred to functionalities provided by the collaborative annotation system as notable positive features of Backstage 2 through statements such as the following:

- “It was really helpful to be able to see the questions and remarks from the other students.”
- “that student is able to work at home but still in "team"/not alone because he can read annotations of other students and the professor. He can add his own as well.”
- “Ask questions that other people can answer”

Other features mentioned positively included that everything required for the course could be found at a single location (3 mentions), the structured representation of the course contents (2 mentions), and the option for synchronizing with the lecturer’s current slide (2 mentions).

To the question *What are for you negative aspects of Backstage?*, 49 students provided an answer. 10 of those mentioned negatively the separation of Backstage 2 into the course delivery platform (which is talked about in this thesis), and the homework submission platform (Backstage 2 / Projects which is reported on in [Hel20]). A few of the responses included criticism of Backstage 2 / Projects itself as well (10 mentions). The most mentioned negative aspect referring to the course delivery platform turned out the be technical problems (12 mentions). 7 of the responses mentioned usability problems but did provide general statements such as “Sometimes not intuitive” or “difficult to use in the beginning”, which did not allow to identify the parts of the system students found cumbersome to use.

Furthermore, responses criticized the collaborative annotation system (8 mentions) and the audience response system (6 mentions). Criticism of the collaborative annotation system encompassed the display of annotations on units and the annotation sidebar in which the textual representation of annotations was displayed, expressed through statements similar to the following:

- “Multiple annotations on the same phrase made it almost impossible to read due to the strong opacity of layered markings”
- “Annotations hide part of the slides”

The audience response system was criticized for an unclear display of quiz results, which was expressed with statements similar to the following:

- "Show solutions of the quizzes more clearly"
- "show, which answers where wrong in a quiz'"

Five responses mentioned negatively the behavior of their peers. Such statements included statements similar to the following:

- "it could have been used more, but that it is a thing of the students."
- "That some questions of the students are still without answers." (Statement was counted as criticism against teaching staff as well)
- "some comments of students were very useless"
- "Some annotations were annoying."

The behavior of the teaching staff was mentioned negatively as well, criticizing the design of the slides and the PDF versions of the slides being uploaded late as well as organizational matters, such as the requirement to create an account for Backstage 2 instead of using students' existing university accounts.

The System Usability Scale measures usability on a scale from 0 to 100 [Bro+96]. Bangor et al. [Ban+09] developed a mapping from ranges on that scale to adjectives. In particular, they assign the range from 50.9 to 71.4 the adjective *OK* but warn that *OK* should not be interpreted as a system's usability being satisfactory and that no improvements are necessary. In both courses, Backstage 2's SUS score laid within that range (63.5 in **LC1**, 58.3 in **LC2**, 59.6 overall).

5.2.4 Discussion

The results from **LC1** and **LC2** generally validate Pohl's result that "Backstage promotes learning-related awareness and activities" [Poh15, p. 81]. Students used the collaborative annotation system during lecture sessions as a backchannel and participated in classroom quizzes. However, both backchannel participation (when considering absolute numbers of annotations), as well as classroom quiz participation (when considering the proportion of active users participating in classroom quizzes) were lower in **LC1** and **LC2** than in the courses examined by Pohl.

Different attitudes and predispositions of the audiences in Pohl's courses compared to the audiences in the courses described in this chapter could be a possible explanation

for that observation: Pohl's courses took place from 2013 to 2015, a time in which technology in large classes was something new which might have evoked a novelty effect in the audiences. The studies described in this chapter took place in 2018 and 2019, respectively, times in which it is more likely for students to have already encountered educational software on earlier occasions, and students, therefore, having already developed a standard for educational software. That standard, however, might not have been fulfilled by Backstage 2. For participation in classroom quizzes, the conservative way of determining users active during a lecture session could have influenced that result: A user was seen as active as soon as a single activity by that user was recorded during the lecture session. Demanding a greater number of activities or continuous interaction might result in more accurate results. Another factor that might have influenced both use of the collaborative annotation system and participation in classroom quizzes is the fact that both courses had their fair share of technical problems: Backstage 2 could only be used actively for six and nine lecture sessions in **LC1** and **LC2**, respectively.

While the results might be worse than the results of Pohl, still, at least half of the students participated in classroom quizzes, and while the number of students participating in backchannel communication during lecture sessions was lower than in the courses evaluated by Pohl, that feature was still used by several students. Moreover, students positively emphasized the functionality and effects of the audience response system and the collaborative annotation system in their answers to the open answer questions.

The lower score for the construct INTERACTIVITY in **LC2** compared to all of Pohl's courses could be explained by the aforementioned different times of the evaluations and Backstage 2 no longer being seen as contemporary. In 2018, and even more in 2019, students are accustomed to nearly anything being possible from their web browser and, therefore, might be less impressed by what is offered by Backstage 2.

Another of Pohl's observations was that students rated high the construct REWORK and that, in two of the examined courses, there was a significant number of students who logged into Backstage outside of lecture sessions. REWORK was rated high by students in **LC1** and **LC2** as well, and the detailed analysis of activity on Backstage 2 showed that students were doing quizzes and using the collaborative annotation system outside the lecture sessions. While the majority of students just read other students' annotations, several students created new annotations or voted on existing annotations outside of lecture sessions. The usage of Backstage 2 followed a pattern that repeated itself each week which suggests that Backstage 2 was used both for preparation and reworking of lecture sessions: The days towards the lecture sessions, activity increased which indicates preparation, and the days after the lecture sessions, activity slowly decreased which indicates reworking. Besides, Backstage

2 was used extensively for examination preparation indicated by a large increase of active users towards the examination. Thus, the results suggest that the quizzes and the units enriched by annotations were seen as valuable learning resources by students. Regardless of the results in **LC1** and **LC2** being worse than Pohl's results, all constructs were still rated positively across both courses.

Backstage 2 was introduced at different times in **LC1** and **LC2**: In **LC1**, Backstage 2 was available to students throughout the term, but only used actively by the lecturer starting from the seventh lecture session. Nonetheless, students still engaged in backchannel communication in the first part. The number of active users drastically increased with the active use of Backstage 2 through the lecturer and only dropped slowly from that point. Backchannel activity in the second part was generally lower than in the first part but still took place, and the majority of active students participated in classroom quizzes. In **LC2**, where Backstage 2 was used actively by the lecturer throughout the whole term (except for the first three lecture sessions), a constant decline of active users throughout the term is visible and less backchannel communication than in the second part of **LC1** took place. Specifically, the number of active users never reached the number of active users in the second part of **LC1**, but from those, the majority participated in classroom quizzes. The observation of the amounts of activity in the second parts of **LC1** and **LC2** combined with significantly fewer students being active in the second part of **LC2** than **LC1** suggests that the belated introduction of Backstage 2 had a reinvigorating effect on the audience. That would mean that not introducing educational software at once, but in parts, could be a way to refresh students' engagement during a term.

Overall, the use of Backstage 2 in **LC1** and **LC2** can be considered a success. While the results of Pohl could not be exactly replicated, the results from the evaluations described here mostly validate his findings. The next section discusses implications for the format Large Class Teaching.

5.3 Wrapping up Large Class Teaching

As Backstage 2 is a prototype only, many areas had to be omitted during the development and for many areas only a rudimentary implementation exists. One area completely omitted is lecturers' awareness, and for communication awareness, only a rudimentary implementation exists.

For a lecturer to effectively teach a large number of students, they have to among others be aware of how their students are doing, if there are questions, and if the lecture session is too fast or too slow. For each of the mentioned areas, different

possibilities are imaginable. The goal, in any case, is to support lecturers in deciding when and what interventions are required.

For an assessment of how their students are doing, Backstage 2 is currently only providing feedback on how students did in a quiz to lecturers. This allows the lecturer to identify areas that were not completely understood and launch an appropriate intervention but does not take into account many other data that is available to the system such as errors students made in their homework submissions and results from previous quizzes. All that data could be aggregated to give lecturers a clearer picture on the understanding of their class and help them to decide how to start a lecture session: Recapitulate contents from the previous lecture session, discuss a common mistake made in the homework submissions, or immediately begin with a new subject.

In Backstage 2, lecturers use the same interface as students to read annotations which makes it difficult for lecturers to identify questions and aspects to address during lecture sessions. In contrast, the previous version of Backstage [Poh15] had more sophisticated means for lecturers' awareness: Lecturers were shown an overview of the distribution of purposes of the created posts which allows them to quickly notice when students start beginning to have problems with the content, for example, by an increasing percentage of questions. That overview could be reset at any time. Furthermore, lecturers were able to filter posts by the number of upvotes so that posts below that threshold were not even displayed and by that, were not cluttering the lecturer's interface.

In another article, Pohl et al. [Poh+12] envision even more powerful filtering options, such as categories, reputations of the posts' authors, or by keyword contained in the posts. Furthermore, they propose that lecture slides can be comprised of regions where a lecture slide can contain more than one region, but regions can span more than one slide as well and allow lecturers to filter annotations by region. As an example, the authors mention a proof that spans more than one slide where all slides referring to the proof have been put into the same region: By that, lecturers can still filter for posts referring to previous pages (and hence, being informed of questions to previous parts of the proof) while talking about the parts of the proof contained on later pages.

Furthermore, the larger the class, the larger the lecture hall and the higher the possibility of lecturers not being near their device that displays the backchannel during the lecture sessions. To notify lecturers on incoming questions, a kind of feedback that is only noticeable to lecturers and independent from their location in the lecture hall is required. Today's smartwatches most often contain a vibration motor which allows, depending on the model, to transmit different vibration patterns

to the wearer's wrist. Such a vibration is unobtrusive, only noticeable to the lecturer, and independent from their location in the lecture hall and allows the lecturer to go back to the device the backchannel is displayed on and decide on an appropriate intervention.

Finally, lecture sessions should not be too fast, but not too slow as well. Thus, lecturers should be aware of how their students are perceiving a lecture session's speed. The first version of Backstage addressed that issue by providing the purposes *too slow* and *too fast* for backchannel posts, which were only rarely used by students [Poh15]. Other backchannel software provides students with buttons for *too slow* and *too fast* (see, e.g., Tweedback [Gar+13]). Another possibility is enabled through the unit-centered nature of Backstage 2: If one presumes that students who have not understood something during the lecture go back to that unit and that students who already understood something leapfrog the lecturer and go to following units, that data can be used to identify how students are perceiving a lecture session's speed. By showing the distribution of students over the units, lecturers can easily notice if the speed of a lecture session is too fast, too slow, or just right: A large number of students lacking behind indicates that the lecture session is too fast; a large number of leapfrogging students that the lecture session is too slow.

There is no doubt that Backstage 2's usability has to be improved indicated by the low score on the System Usability Scale and several users listing various usability issues in the free text questions of the surveys. One potential source of usability issues are devices students bring with them to lecture sessions: While at the beginning of the decade, those devices were mostly laptops, nowadays, students most often bring tablets or smartphones to lecture sessions, for which Backstage 2 is in no way optimized. Therefore, one way to improve usability might be to make Backstage 2 responsive, so that the majority of functionality can be used across various types of devices. Furthermore, dedicated usability studies to identify areas that are lacking usability should be conducted. The surveys used in the evaluations described in this chapter had usability only as an afterthought, and there is no question that dedicated usability surveys would be more effective in unearthing usability issues of Backstage 2.

Similar to Pohl, the evaluations in this chapter did not measure learning, but whether Backstage 2 promotes learning-related activities in students. To measure learning, different groups receiving different treatments are required, which are difficult to obtain when studies are done in real teaching contexts. A comparison between groups is often done using grades or by administering pre- and post-test surveys. Another possibility would be to use an experimental design in real teaching scenarios and apply only a single change between two venues. Nonetheless, applying only

a single change just for the sake of an experiment even if there are other possible improvements is ethically highly questionable.

To improve how students perceive INTERACTIVITY on Backstage 2, more of the functionalities of Backstage 2's audience response system, in particular, its ability to support a variety of possibly more engaging question types, could be utilized. These question types could be introduced bit by bit, and so could maybe leverage the reinvigorating effect described in the previous section.

When teaching a large class using the format described in this chapter, the lecturer is still orchestrating most of the class interaction. In the next format, *Phased Classroom Instruction*, lecturers take a step back and mainly support students while they work on larger exercises during classroom sessions.

6 Phased Classroom Instruction

This chapter is based on the following articles:

- Sebastian Mader and François Bry. "Phased Classroom Instruction: A Case Study on Teaching Programming Languages". In: *Proceedings of the 10th International Conference on Computer Supported Education*. SciTePress, 2019, pp. 241–251
- Sebastian Mader and François Bry. "Promoting Active Participation in Large Programming Classes". In: *Computers Supported Education*. Springer, 2020, to appear

In addition to the contents of the article, this chapter includes with **PCI3** a further evaluation, and accordingly, a description of the changes made for that venue.

While the previous chapter demonstrated in which ways Backstage 2 can make traditional lectures more student-centered, the use itself is still mostly orchestrated by a lecturer: Except for the communication that takes place in the backchannel, every other interaction is initiated by a lecturer. The format Phased Classroom Instruction, described in this chapter makes more room for student-centered work with the lecturer transitioning from "[f]rom [s]age on the [s]tage to [g]uide on the [s]ide" [Kin93, p. 30]. A typical lecture session in Phased Classroom Instruction consists of several blocks each comprising of a mini-lecture, a practical exercise in which the concepts taught the in mini-lecture are immediately applied, and a peer review in which each student is assigned another students' submission for review.

Phased Classroom Instruction is strongly related to the learning and teaching format *flipped classroom*, where "events that have traditionally taken place inside the classroom now take place outside the classroom and vice versa" [Lag+00, p. 32], that is, students work on practical exercises during lecture sessions applying the knowledge they acquired outside the classroom using lecture material provided by lecturers, most often in the form of videos [Gia+14]. During the exercises, the lecturer stands ready to assist the students who require support in solving the exercises [PK13].

Flipped classrooms face various issues, which are discussed in detail in Section 6.1: For one, much effort is needed for the creation of learning material for self-study, as such materials generally have to be more bullet-proof than traditional lecture slides with no lecturer at hand to provide clarifications and corrections. Furthermore, students often have to be incentivized with external rewards (most often course credit) for them to actually engage with the learning material outside of

class. Engagement with the learning material before class is a cornerstone of flipped classrooms as it is a precondition for the practical exercises. Furthermore, flipped classrooms do not scale particularly well to larger classes as there is an upper bound to the number of students a lecturer can effectively support during the exercises. The examined implementations of flipped classrooms in Section 6.1 suggest that this upper bound lies somewhere between 20 and 40 students.

The aforementioned issues are addressed in two ways in Phased Classroom Instruction: Instead of students learning the subject matter on their own, lecturers address the subject matter in a short mini-lecture at the beginning of a block of Phased Classroom Instruction. That mini-lecture is followed immediately by an extensive exercise in which students put the just acquired concept to use. Technology is used to scale Phased Classroom Instruction to large classes: Students work on the exercises using problem- or subject-specific editors, which provide students immediate feedback and scaffolding (see Chapter 4) enabling those students who just require a nudge in the right direction to solve the exercise without a lecturer's personal support. That, in turn, frees up time for lecturers to focus on those students that require their personal support. To identify those students, technology provides lecturers an overview of the class which allows lecturers to identify whom to help without having to walk through the lecture hall glancing at screens and papers to get the same overview. As a final phase, Phased Classroom Instruction includes a peer review, where each student is assigned another student's submission for review.

Phased Classroom Instruction uses Backstage 2's audience response system with the already discussed three-phase quiz (see Section 4.2.3) in which students first create a submission using a problem- or subject-specific editor and are then provided another student's submission for review. In the case of the evaluations described in this chapter, a subject-specific editor for the programming language JavaScript was used.

Phased Classroom Instruction is similar to Frederick's proposal of "[a]lternating [m]ini-[l]ectures and [d]iscussions" [Fre86, p. 47] with the difference that discussions are replaced by exercises and that Frederick's proposal is intended to be used without supporting technology.

This chapter is structured as follows: The next section reviews implementations of flipped classrooms and discusses the issues which were shortly broached in this introduction. After that, Phased Classroom Instruction is introduced in more detail. Then, the first two venues Phased Classroom Instruction was evaluated in and the technological supported used are introduced, before the results of the evaluations of those venues are presented and discussed which point to the format being well-liked by students, but reveal weaknesses as well. After that, adaptions made to the

technological support and the course in response to the uncovered weaknesses are described, and results of an evaluation with the adaptions in place in a third venue are presented and discussed which point towards the adaptions being successful. The last section summarizes the chapter and gives an outlook at further research perspectives for Phased Classroom Instruction.

6.1 Flipped Classrooms

Flipped or *inverted* classrooms are a learning and teaching format that, driven by the technological advances during that time [Bak00; Lag+00], began to emerge at the beginning of the century. Lage et al. [Lag+00] *invert the classroom* to be able to employ a wider variety of teaching styles to cater to the variety of learning styles of students. According to them, "[i]nverting the classroom means that events that have traditionally taken place *inside* the classroom now take place *outside* the classroom and vice versa" [Lag+00, p. 32]. At the same time, Baker [Bak00] introduced his idea of a *classroom flip*, which is the "movement of lecture material out of the classroom through online delivery" [Bak00, p. 12], which enables lecturers to "use class time for other activities" [Bak00, p. 13] and by that introduce active learning to lectures. In more recent literature, Bishop and Verleger [BV+13] see flipped classrooms as "[consisting] of two parts: interactive group learning activities inside the classroom, and direct computer-based individual instruction outside the classroom" [BV+13, p. 4] and restrict their definition further to only include formats that use videos as learning material outside the classroom. Their definition is more in line with the understanding of Baker [Bak00], who saw flipped classrooms as an opportunity to introduce active learning to lectures. In summary, the main idea of flipped classrooms is that students acquire the subject matter outside of the classroom and apply the subject matter in practical exercises during lecture sessions.

Flipped classrooms have been evaluated on a wide range of subjects, such as statistics [Wil13; Str12; Tal13], biology [Sto12], business administration [Sch+11], economics [Lag+00], pharmaceutics [McL+14], physics [Ste+10], nutrition studies [Gil+15], engineering [Bla+16], and computer science [RB15; Cam+14; Gan+08; LE13; GPI13; Amr+13; Mah+15; Sar14; KF05].

Immediate feedback and the ability to address misconceptions as they arise are one of the strengths of flipped classrooms [Lag+00], but to do so appropriately the lecturer requires an overview of their class. While Lage et al. [Lag+00] see room for a few more students in their flipped classroom of 40 students, for larger classes they propose to address the issue by adding more teaching staff to a lecture hall or breaking the lecture down into smaller sections. Similar sentiments are brought up

Tab. 6.1.: Overview of the class sizes in various implementations of flipped classrooms.

Study	Class size
Reza and Ijaz Baig [RB15]	23
Schullery et al. [Sch+11]	24
Stelzer et al. [Ste+10]	24
Wilson [Wil13]	20 – 25
Sarawagi [Sar14]	26
Gannod et al. [Gan+08]	24, 22, 27[1]
Strayer [Str12]	27
Herold et al. [Her+12]	36 – 38
Gilboy et al. [Gil+15]	24, 37[2]
Lage et al. [Lag+00]	40
Lockwood and Esselstein [LE13]	30 – 40[3]
Gehringer and Peddycord III [GPI13]	8, 44
Blair et al. [Bla+16]	42
Maher et al. [Mah+15]	11 – 93[4]
McLaughling et al. [McL+14]	162
Campbell et al. [Cam+14]	190[5]
Stone [Sto12]	30, 400

[1] 24 students in one section, 43 students in two sections, 80 students in three sections; values obtained by assuming sections of equal size

[2] 148 students in four sections, 48 students 2 sections; values obtained by assuming sections of equal size

[3] "The course meets in a computer lab that seats 30 students, but is often over-subscribed with at least 35 students enrolled in each section." [LE13, p. 114]

[4] unclear, if students are split into sections or not; in [Mah+13] one of the larger courses is described in more detail using sections

[5] 570 students in three sections; values obtained by assuming sections of equal size

in more recent literature as well [Cam+14; Sar14; Gan+08]. Indeed, the majority of studies on flipped classrooms evaluated the format in classes of 40 students as less, as can be seen in Table 6.1.

In four of the studies described in Table 6.1 flipped classrooms were evaluated in larger classes. The following paragraphs shortly discuss those studies and propose reasons why those studies might have succeeded to deploy the format in large classes.

In the study done by Maher et al. [Mah+15] it is unclear whether the courses were taught in sections; in an earlier article, Maher et al. [Mah+13] present the evaluation of one of the courses also mentioned in the later article which indicates that the course was taught in sections (of 45 students each) in one year, and in the following year in one section of 90 students for quiz activities and two sections

of 45 students for programming labs. Hence, it can be assumed that the authors noticed that some forms of active learning (in that study, activities in programming labs compared to quiz activities) require more lecturer intervention than others and should, therefore, be taught in smaller groups.

Using forms of active learning that require less lecturer intervention might be the reason why McLaughling et al. [McL+14] were able to flip a class of 162 students. These authors mainly used activities such as quizzes using an audience response system, student presentations, or pair and share activities. Campbell et al. [Cam+14] applied their flipped classroom to course sections of around 190 students, but, on average, only 57% of all students attended the lectures where the lecturer was supported by 1 to 2 teaching assistants, which results in around 36 to 54 students per teaching staff.

In Stone's [Sto12] course with 400 enrolled students, on average, 80% of the students attended the lectures in which activities that would normally require more support through a lecturer, such as discussing and working on past examination questions and problem-solving, were done. Stone unfortunately does not shed light on how the lecturer was able to support such a large number of students during the classroom sessions.

Regarding the materials provided to students for self-learning outside of lecture sessions, most implementations of flipped classroom use videos [Gia+14], with a few implementations including quizzes inside their videos [Cam+14; Ste+10] or providing simulations [Sto12]. One issue associated with these self-learning materials is the significant effort associated with their creation [Tal13; Gan+08; Gia+14; Sar14; Gil+15; LE13; HS13]. A few articles provide concrete figures on the time taken for the creation of the materials: Kaner and Fiedler [KF05] mention 7.5 to 25 hours required for one hour of final video, Campbell et al. [Cam+14] quote 600 hours required for the whole material of their course, Stelzer et al. [Ste+10] required around 1400 hours (28 units each taking 50 hours) to complete their course, and Maher et al. [Mah+15] mention 300 hours (distributed over three terms) for 26 videos.

Another issue with students learning the subject matter outside class is getting them to actually view and work with the material. Often instructors attached course credit to get students to actually engage with the lecture material outside of lecture sessions: Course credits were awarded for successfully completing quizzes about the material either before lectures [Gil+15; GPI13; LE13; Cam+14] or during lectures [Sar14; RB15].

Listing the various types of activities done during the lectures is out of the scope of this thesis, but the activities were often done in pairs or teams and *active* in the sense that students were more involved than just listening to the lecturer (see, e.g., [Sto12; KF05; Cam+14]). The time effort for preparing the activities that are done during lecture sessions is mentioned as a downside of flipped classrooms as well [Gan+08; Gil+15; Gia+14; Sar14]. Campbell et al. [Cam+14] mention taking 130 hours for the creation of the in-class exercises.

Some of the implementations of flipped classrooms include a lecture-esque component, either pre-planned [Wil13] or adaptively as an intervention to address misconceptions and confusion as they arise [McL+14; Mah+15].

Regarding the effects of flipped classrooms, the majority of articles report on positive attitudes of students towards the format [Sto12; Cam+14; McL+14]. Few studies report on the effects of flipped classrooms on students' performances: While some of those studies report on a positive influence on students' performances [Sto12; RB15; Wil13], there are studies which find no effect [Cam+14; GPI13; Bla+16]. However, none of the studies reports on a negative effect on students' performances.

A similar (mixed) pattern can be seen regarding the effects of flipped classrooms on attendance: Some studies found that more students attend flipped classroom lectures compared to traditional lectures [RB15; GPI13; Sto12; Ste+10], but other studies found the opposite [Cam+14; Bla+16; Bla+16].

Even though the effects on students' performances are unclear, flipped classrooms are a format students are enjoying very much, which makes it, as long as there is no evidence for negative effects, a suitable approach for making classes more interactive and engaging.

The remainder of this chapter introduces Phased Classroom Instruction in more detail and how the format addresses the issues of flipped classrooms before introducing the evaluations of the format in three courses on software development.

6.2 Phased Classroom Instruction

A block of Phased Classroom Instruction consists of three phases:

1. A *mini-lecture* of about 10 to 20 minutes in which a concept is introduced.
2. An extensive *exercise* in which students put the just acquired concept to work either in individual work or teamwork.

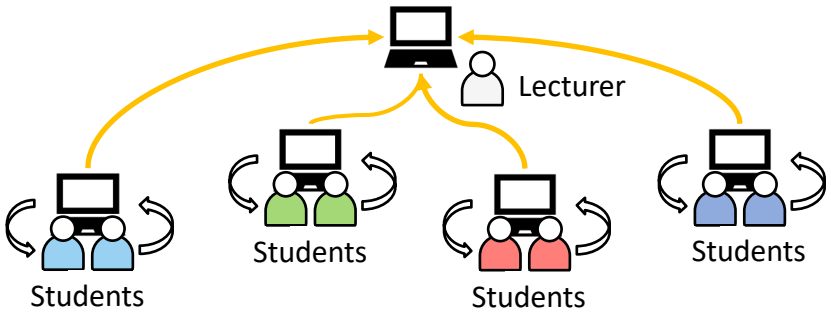


Fig. 6.1.: Schematic overview of the role of the technological support in Phased Classroom Instruction.

3. A *peer review* in which each student is assigned another student's submission for review.

A lecture session may consist of one or more blocks of Phased Classroom Instruction depending on the duration of the lecture sessions and the extent of the exercises. Putting a lecture in form of a mini-lecture back into lecture sessions reduces the effort for lecturers, as course material no longer has to be as bullet-proof as for self-learning and time-extensive production of videos is no longer necessary. Additionally, the mini-lecture addresses the problem of students not engaging with the learning materials outside the classroom.

The limited scalability of flipped classrooms is addressed by Phased Classroom Instruction by providing technological support for students and lecturers alike. Figure 6.1 shows schematically how technology enables Phased Classroom Instruction to scale to larger classes.

In the bottom part of that figure, students are working on the exercises using a problem- or subject-specific editor which supports them with immediate feedback and scaffolding while working on the exercise. The support provided by those editors alone might be enough for those students who are "nearly there" to solve the exercise successfully without a lecturer's support. While students are working on the exercises, the current submissions of all teams are automatically analyzed to provide lecturers an overview of the class to help them to identify whom to help. Analyzing can either be done using exercise-specific measures (e.g., number of passing unit tests), exercise-independent measures (e.g., time idle), or a mix of both. Phased Classroom Instruction concludes with a peer review, where each student is assigned another student's submission for review.

The revised version of Bloom's Taxonomy of Learning Objectives (see [Blo56]) by Krathwohl [Kra02] consists of the levels "Remember", "Understand", "Apply",

"Analyze", "Evaluate", and "Create" [Kra02, p. 215] where each level depends on the levels below it and is meant as a means for classifying the goals of teaching. There are more aspects to the taxonomy which are omitted here as the levels are sufficient for understanding the following; refer to [Kra02] for a detailed overview of the taxonomy. A few of the studies on flipped classrooms argue that the outside class activities are attending to the lower two levels, "remember" and "understand", of the taxonomy, while the in-class activities attend to the levels above those [Gil+15; Sar14]. A similar argument can be made for Phased Classroom Instruction: The mini-lecture covers (analogous to flipped classrooms) "remember" and "understand", the exercise "apply" and "analyze", and the peer review "evaluate". The final step of the taxonomy, "create", is hard to implement as the type of exercises that can be conducted in a lecture session will always be somewhat constrained due to time reasons.

Phased Classroom Instruction was evaluated in three courses on JavaScript programming. The following section describes the evaluations of the first two courses.

6.3 First Steps with Phased Classroom Instruction

Phased Classroom Instruction was evaluated in the lecture sessions accompanying a practical on game development using the programming language JavaScript. In the lecture sessions, which take place during the first weeks of the term, students learn the basic concepts of the programming language and game programming, before they start to implement a larger project in teams of four students.

This section first describes the technological support for students and lecturers and introduces the two venues of the course in which Phased Classroom Instruction was evaluated in more detail before presenting and discussing the results of the evaluations.

6.3.1 Technological Support in the first two Venues

As already mentioned, Phased Classroom Instruction was implemented in Backstage 2 using its audience response system and the three-phase quiz which lets students first work on an exercise using a JavaScript editor, and afterward presents each student another student's submission for review. Finally, each student is shown their submission and the review. Furthermore, a new type of unit was implemented for Phased Classroom Instruction: Scaffolded exercises that present an exercise subtask by subtask. The following section introduces these technological supports which were used for the first two venues.

The First Version of the JavaScript Editor

JavaScript is a programming language that runs directly in web browsers, and JavaScript code can be added to websites. Adding JavaScript code to websites allows among others to add, delete, or update elements on the website as well as draw various geometric primitives on a so-called `canvas` element. That `canvas` element is the main element used in the software development practical, as the whole game developed in the practical is drawn onto such an element. The following section describes the first version of the JavaScript editor which was developed by Maximilian Meyer [Mey19] as part of his master thesis. Refer to his master thesis for a more detailed description and an evaluation of the editor.

The use of the `canvas` element in the software development practical made it a requirement for the editor to support visual output as well, that is, displaying what JavaScript code draws on a `canvas`. Visual output was implemented using an `iframe` element which contains a `canvas`. Executing the code in that `iframe` shows the result of the execution on the contained `canvas`. Moreover, the `iframe` acts as a security layer, as code executed in an `iframe` cannot affect the website it is embedded in.

A screenshot of the JavaScript editor used by students to create their submissions can be seen in Figure 6.2: The tabs above the text area allow to switch between various modes of the editor. The tabs *JavaScript* (the currently selected tab) and *HTML* show text areas in which code in the respective language can be entered. While typing in those text areas, basic error messages on syntax errors (e.g., when a closing or opening bracket is missing) and indentation support (e.g., pressing enter after defining a function automatically indents the following line by 4 spaces) are provided. The play button on the top right executes the code in the `iframe`.

The result of that execution can be examined in the tab *Output*. Figure 6.3 shows the output that is generated by running the code shown in the editor in Figure 6.2.

The tab *Console* shows a prompt in which JavaScript expressions can be evaluated and displays calls to `console.log`. Furthermore, the console allows users to interact with the current state of the executed code, that is, users can call functions defined in the code, check the values of variables, or modify the values of variables.

The last tab, *Testing* shows all unit tests that are defined for the current exercise which are run each time the code is executed. Figure 6.4 shows an example of the content of the Testing tab, which shows two passing and one failing test. Each test

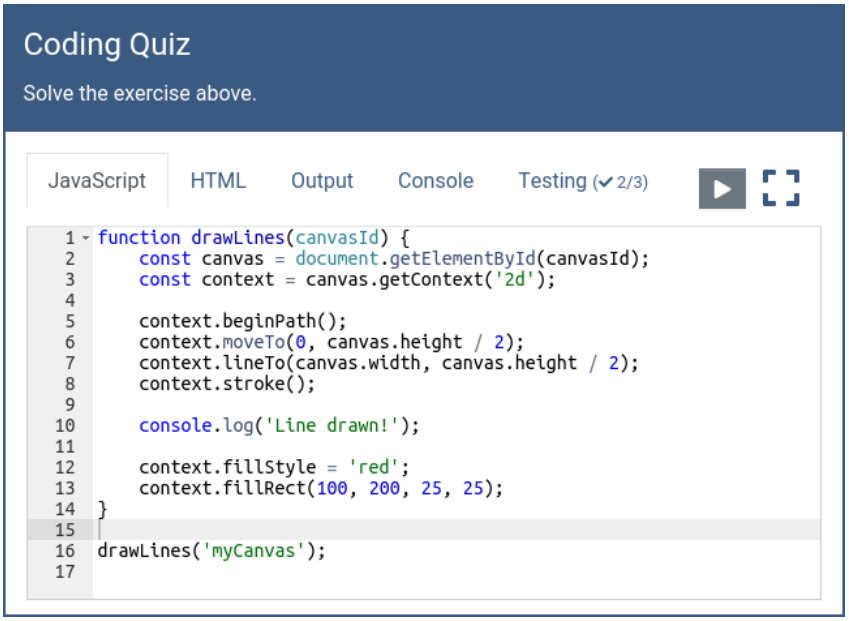


Fig. 6.2.: Screenshot of the web-based JavaScript editor used in **PCI1** and **PCI2**.

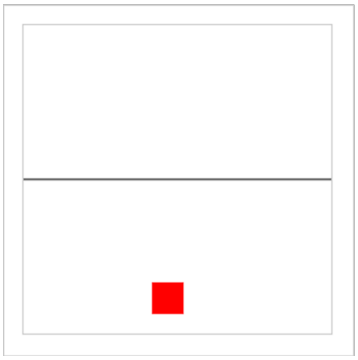


Fig. 6.3.: Result of executing the code shown in Figure 6.2.

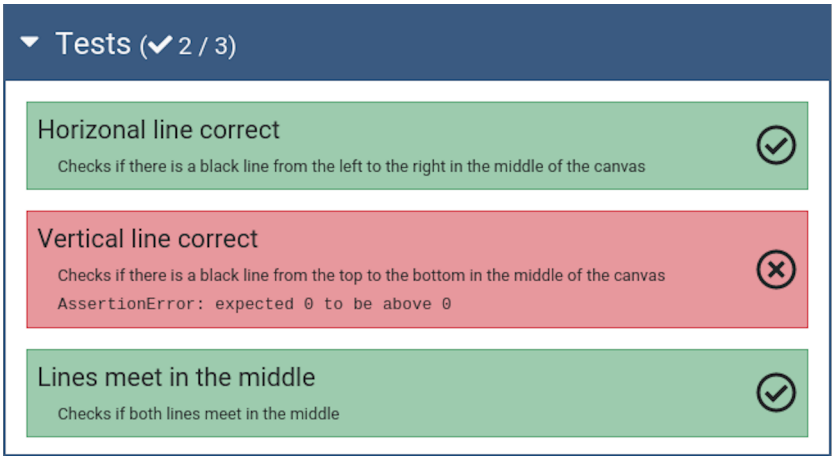


Fig. 6.4.: The *Testing* tab of the JavaScript editor after executing the code resulting in two passing and one failing test. For failing tests, the error message returned by the testing framework is displayed below the description.

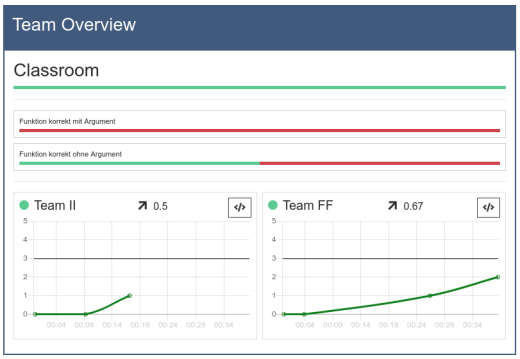


Fig. 6.5.: Classroom overview showing for each team the number of passing tests over time and the slope of that graph. The bars in the middle show all tests that at least one team is failing (taken from [MB20, p. 11]).

consists of a title, a description of what is being tested, and, in case of failure, the error message returned by the used testing framework.

Class Overview

The class overview introduced in the following replaces the overview of quiz answers usually shown while a quiz is running (see Section 4.3). The class overview uses the number of passing unit tests over time as the measure to allow lecturers to identify struggling students. The underlying intuition is that more successful teams can pass more unit tests in shorter time and vice versa. A screenshot of the class overview can be seen in Figure 6.5.

In that figure, each team is represented by a graph which shows the course of passing unit tests over time. Additionally, the slope of that graph is shown for each team

(see the number next to each team name). A team's current code can be accessed by clicking the button on the right of each team. Above the individual teams' results, all unit tests that are failed by at least a single team (and the distribution of teams failing and passing that test) are shown to help the lecturer to identify general problems of the students.

Interface for Scaffolded Exercises

Recall that quizzes in Backstage 2 are always attached to a unit. For representing the exercises' tasks, a new type of unit was implemented that presents the subtasks of an exercise subtask by subtask. Each exercise was divided into several subtasks, and the unit did not show all subtasks at once but rather demanded from students to manually unlock the next subtasks by clicking a button after the current subtask has been completed. The rationale behind that design is to not overwhelm students with a large number of subtasks but to help students focussing on one subtask at a time.

The JavaScript editor, the class overview, and the interface for displaying scaffolded exercises were used for the evaluation of Phased Classroom Instruction in the first two venues. Note that not every part was used in both venues. The next section details which parts were available in which venue and reports on the evaluations of Phased Classroom Instruction in those venues.

6.3.2 Study

The remainder of this section is dedicated to the evaluations of Phased Classroom in the first two venues of the software development practical, **PCI1** and **PCI2**. Various factors contribute to the success of a format such as Phased Classroom Instruction: Students have to be able to solve the exercises in the allotted time as not being able to solve the exercises might demotivate them. Additionally, students should be able to finish the exercises in similar amounts of time, as finished students might disturb students still working on the exercise. Finally, as for every learning format, the students' attitude might be the most important factor, as without students liking the learning format, teaching gets difficult. Note that the three venues were taught by the author of this thesis. Hence, references to a *lecturer* refer to the author of this thesis.

The first two venues of the software development practical took part during winter term 2018/19 (**PCI1**) and summer term 2019 (**PCI2**) and had several differences. An overview of the differences between the two venues can be seen in Table 6.2.

Tab. 6.2.: Overview of the differences between **PCI1** and **PCI2**.

	PCI1	PCI2
# of participants	16	44
# of lecture sessions	5	6
duration of lecture sessions	125 minutes	90 minutes
# of phased classroom blocks	11	9
# of unit tests available	0	5
lecturer overview used	no	yes
exercises displayed scaffolded	no	yes
peer review conducted	yes	no

As **PCI1** was thought of as a first test run of the technological support and the course material, only 16 students were admitted. In this venue, the mini-lecture and the exercises were structured around the video game Snake.[1] Every mini-lecture addressed one concept of game development, such as moving game elements or drawing assets, and students applied the taught concept immediately in the following exercise to their own, ever-evolving version of Snake. Through the exercises, every team should have implemented their own version of Snake at the end of the last lecture session. Only a few of the exercises were structured as described above (consisting of a general description and subtasks), with the majority just being a more detailed general description. Moreover, due to time constraints, no unit tests were available in that venue. Consequently, the class overview was not used in **PCI1** as it depends on the results of unit tests.

Unfortunately, in **PCI2**, organizational matters forced the decrease of the duration of the lecture sessions by 45 minutes to 90 minutes. That change made an implementation of Snake as well as conducting peer review no longer viable. Hence, the extent of the exercises was reduced whereas the idea of exercises building upon each other and working on an ever-evolving piece of code was retained: Instead of implementing Snake, students still applied the just taught concept and implemented a square, made it moveable using the arrow keys, and finally replaced the square with a graphic. Furthermore, all exercises were phrased in the aforementioned way of a general description followed by several subtasks, and unit tests were written for five of the nine exercises. For displaying the exercises' tasks, the aforementioned unit which shows these subtask by subtask was used. As unit tests were available for some of the exercises, the classroom overview was used during the working on those exercises.

[1] `https://en.wikipedia.org/wiki/Snake_(video_game_genre)`

Methods

The following section is a slightly adapted reproduction as found in [MB20, p. 14f.].

Both courses were evaluated using the same survey, therefore the description of the survey is a verbatim reproduction as found in [MB19c, p. 247]. The remaining paragraphs were revised or added to reflect changes in the evaluation.

Data for the evaluation was collected using a survey and taken directly from Backstage 2's database as well. The surveys were conducted during the final lectures of each course and consisted of the following six parts. The entire survey can be found in Appendix A.2.

1. Four questions referring to the students' course of study, current semester, gender, and team they were in.
2. Six questions measuring the students' attitude towards the course format and its elements.
3. Six questions measuring the students' attitude towards the content and structure of mini-lectures and exercises.
4. Six questions measuring the students' attitude towards the enabling technology.
5. Five questions measuring the students' programming proficiency using an adapted version of the survey by Feigenspan et al. [Fei+12].
6. Three questions in form of free text questions, asking about what they liked most, what could be done better, and for further comments.

For parts (2), (3), and (4), a six-point Likert scale from *strongly agree* to *strongly disagree* with no neutral choice was utilized. In the reported results below, *strongly agree* was assigned the value 5; *strongly disagree* the value 0.

All submissions were retrieved directly from Backstage 2's database. A single lecturer determined for each team and exercise the point in time in which the exercise – if at all – was solved correctly. The correctness of an exercise was determined strictly: A submission was seen as correct, if and only if the whole task was solved correctly. That means that nearly correct submissions (e.g., a rectangle moving into the correct direction for three of the four arrow keys) were classified as wrong.

Due to internet connectivity problems in **PCI1**, data for the first lecture is not complete, as not all teams were able to connect to the platform and is, therefore, omitted from the evaluation.

Tab. 6.3.: Overview of the population of **PCI1** and **PCI2**.

	PCI1	PCI2
# of registered students	16	44
# of survey participants	16	32
Average coding proficiency	3.5	3.7

Significance was determined using the Mann-Whitney U test, as the majority of data does not follow a normal distribution which calls for a non-parametric test (see [CF14]). The significance threshold was set to $p = 0.05$. Aggregated measures are reported as Median, hereafter abbreviated as Mdn, as it is more robust against outliers [How09], therefore deviation is reported as Median Absolute Deviation, hereafter abbreviated as MAD (see [RC93]).

Results

Table 6.3 shows an overview of the number of course and survey participants, as well as the average coding proficiency of the surveys' participants. In both venues, the majority of students took part in the survey, but as the surveys were conducted during the last lecture session, no opinions of students not being present during that session are available. In both courses, students exhibited a similar coding proficiency, with the Mann-Whitney U test indicating no significant difference ($p = 0.3$).

This section considers first how successful the teams were in solving the exercises during the lecture sessions, presents numbers on students' attendance during the lecture sessions, and then reports on the students' attitudes and opinions on Phased Classroom Instruction. This section is closed by reporting on the lecturer's personal experiences made while teaching.

Exercise Correctness and Working Time Which exercises were solved correctly (or not) by which teams during the lecture sessions can be seen in Figures 6.6 and 6.7 for **PCI1** and **PCI2**, respectively. The first digit of each exercise is the lecture session in which that exercise was worked on, and the second digit is the number of the exercise within the lecture session. The trailing number of each row indicates the percentage of exercises solved correctly by the respective team, and the last number of each column represents the percentage of teams solving the respective exercise correctly during the lecture session.

In both venues, teams were similarly successful in correctly solving the exercises during the lecture sessions (Mdn: 43.8%, MAD: 9.3 for **PCI1**, Mdn: 44.4%, MAD:

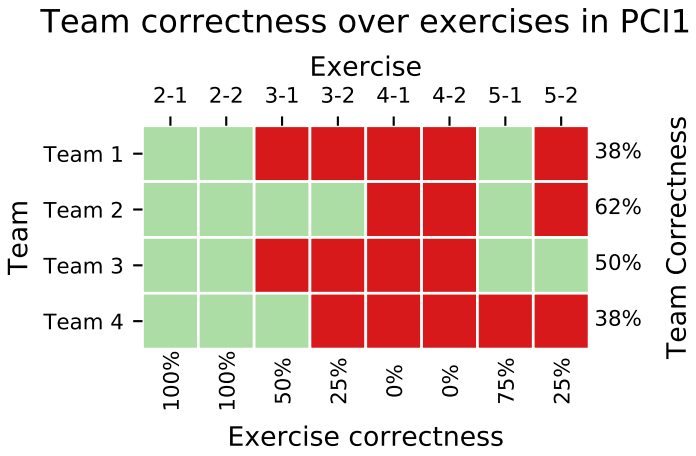


Fig. 6.6.: Overview of exercises being solved correctly during lecture sessions by team and exercise for **PCI1**. A green square indicates that the team was able to solve the exercise correctly during the lecture session; a red square that the team was not able to solve the exercise during the lecture session.

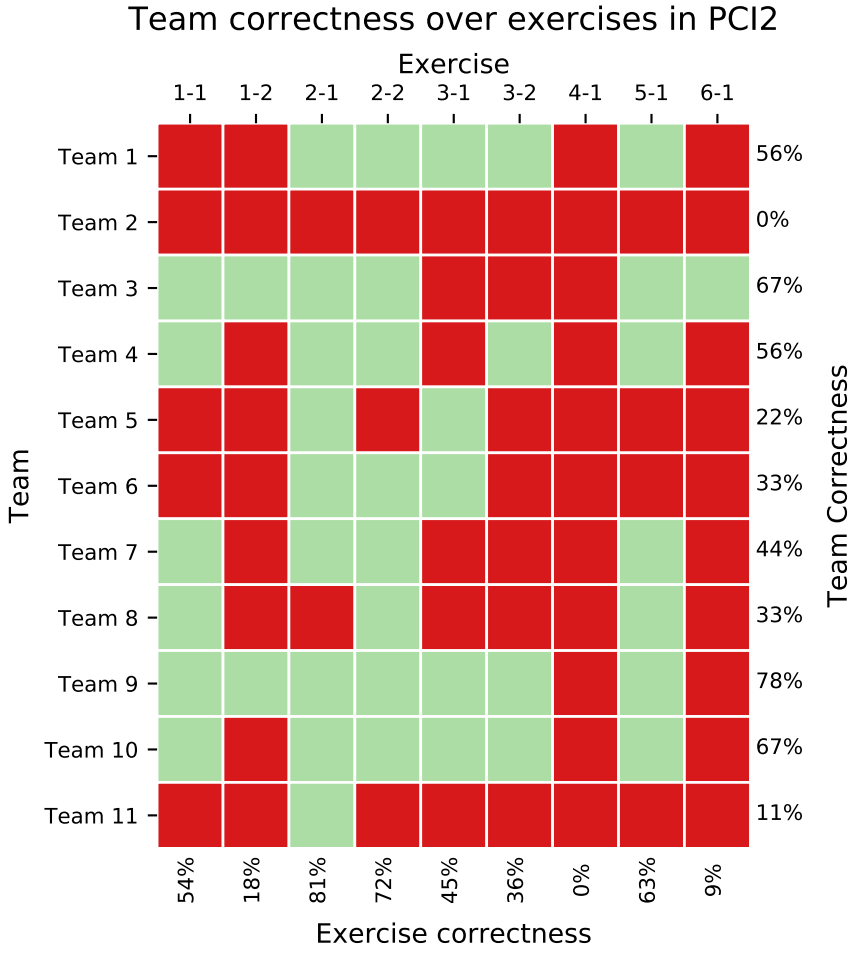


Fig. 6.7.: Overview of exercises being solved correctly during lecture sessions by team and exercise for **PCI2**. A green square indicates that the team was able to solve the exercise correctly during the lecture session; a red square that the team was not able to solve the exercise during the lecture session.

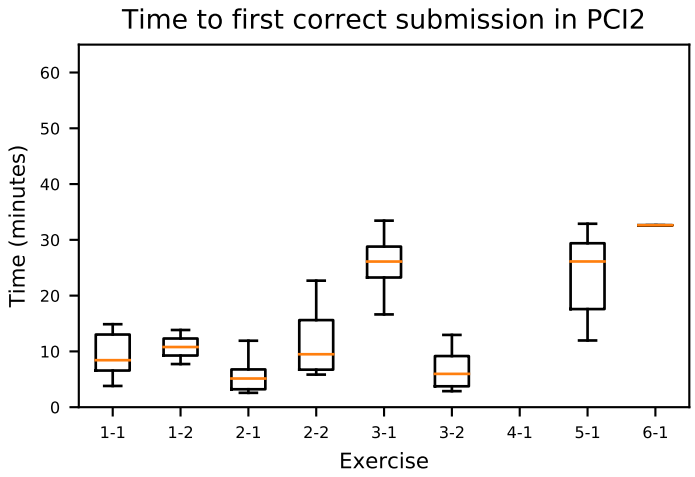


Fig. 6.8.: Time to first correct submission by exercises for **PCI2**.

33.0 for **PCI2**). Teams generally failed to solve even half of the exercises correctly during the lecture sessions, but there were few teams in both venues that solved at least half of the exercises correctly during lecture sessions, such as Team 2 in **PCI1** or Team 3 in **PCI3**.

Discounting the first two exercises in **PCI2** (as results for those are most likely incomplete due to software problems), teams were generally more successful in earlier exercises than in later exercises, that is, with increasing difficulty, teams were less likely to solve the exercises correctly. Especially exercise *4-1*, the exercise on object-oriented programming, was in both courses solved by none of the teams. In **PCI2**, *4-1* is the only exercise solved by none of the teams, in **PCI1**, *4-2* is solved by no team as well. The bad performance of teams in **PCI1** in exercise *4-2* is a direct effect of their performance in *4-1*, as when the lecturer saw teams struggling with *4-1*, more time was given for working on that exercise, leaving only a little time for *4-2*.

Figure 6.8 shows the time teams took for their first correct submission in **PCI2**. Due to the low number of teams and correct submissions in **PCI1**, the figure is omitted for that venue. Note that as this figure excludes teams that never turned in a correct submission, when interpreting the graph, the percentage of teams actually solving the exercise has to be taken into account: Exercises *2-1* and *2-2*, which were solved successfully by the majority of teams, show that both extremes exist in the working times: Teams finishing in close succession in *2-1*, but large differences between teams in *2-2* as well.

Tab. 6.4.: Percentage of students being present during lecture sessions as counted by the lecturer in **PCI2**.

Course	L1	L2	L3	L4	L5	L6
PCI2	100%	98%	91%	70%	70%	73%

In median, the interquartile range, that is, the difference between the first and third quartile, or the height of the bars in Figure 6.8, is 5.5 minutes (MAD: 3.2). The highest difference was 11.8 minutes, the lowest 3.1 minutes.

Summarizing the results from this paragraph, teams were generally unsuccessful in solving the exercises with the majority of teams failing to solve half of the exercises. Times until correct submission varied between teams depending on the exercise.

Attendance during Lecture Sessions While in **PCI1**, the lecturer did not specifically count the number of students being present, in most cases, all teams were complete with only one or two students missing on a few occasions. In **PCI2**, students attending each lecture session were counted by the lecturer and can be seen in Table 6.4. Attendance in **PCI2** was above 90% for the first three sessions and then dropped, but never below 70%.

Overall, the attendance in both venues was high with, in any case, more than 70% of all students being present during the lecture sessions.

Students' Attitude towards Phased Classroom Instruction As mentioned in Section 6.3.2, the survey contained among others three blocks which measured the students' attitudes towards Phased Classroom Instruction, the course material, and the technological support. This paragraph introduces the aggregated student responses to these blocks.

The responses for students' attitudes towards Phased Classroom Instruction can be seen in Table 6.5. Across both courses, students viewed the format and its components favorable: Students found the exercises and discussions within the teams helping them understand the subject matter and strongly disagreed with the statement of preferring a traditional lecture to Phased Classroom Instruction.

The results to the questions about the students' attitudes towards the technological support, which can be seen in Table 6.6, show a similar picture across both venues: Students found Backstage 2 to support Phased Classroom Instruction well. While students found that the web-based JavaScript editor helped them getting started with JavaScript and was easy to operate, the majority of students would have rather

Tab. 6.5.: Results of the survey block measuring the students' attitude towards Phased Classroom Instruction for **PCI1** and **PCI2**.

Statement	PCI1 Mdn	PCI2 Mdn
The immediate practical exercises after the mini lectures helped me understand the topic.	5.0	5.0
Discussions with my team mates during the practical exercises helped me understand the topic.	5.0	4.0
I would have preferred a traditional lecture without practical exercises.	0.0	1.0
I had fun during the plenum sessions.	4.5	4.0

Tab. 6.6.: Results of the survey block measuring the students' attitude towards Backstage 2 for **PCI1** and **PCI2**.

Statement	PCI1 Mdn	PCI2 Mdn
The JavaScript editor on Backstage made the getting started with JavaScript easy.	3.5	4.0
The JavaScript editor was easy to operate.	3.5	4.0
The interface of Backstage, where exercises were worked on, was clearly designed.	4.0	4.0
The course format (i.e., mini lectures, followed by exercises and peer review) was well-supported by Backstage.	4.0	4.0
I would have preferred to solve the practical exercises using a real development environment.	3.0	3.0

used a real development environment, that is, software running on their computer, for solving the practical exercises.

To the questions measuring the students' attitude towards the course material, students, again, show nearly identical attitudes across both venues. Students liked exercises building upon one another and found the exercises neither too difficult nor too extensive. According to the students, the mini-lectures were sufficient to solve the exercises. The detailed results to those questions can be seen in Table 6.7.

Summing up the students' attitudes which were nearly identical across both venues: Students liked the format and found its components helpful. Backstage 2 supported the format well, but students would have preferred to work on the exercises using a real development environment. The exercises were of an appropriate difficulty and extent, and students liked that the exercises built upon each other.

Students' Opinions The following paragraph presents common themes in the students' answers to the free text questions. Students' answers were not divided by

Tab. 6.7.: Results of the survey block measuring the students' attitude towards the course material for **PCI1** and **PCI2**.

Statement	PCI1 Mdn	PCI2 Mdn
The mini lectures were sufficient to solve the practical exercises.	4.0	4.0
I would have preferred exercises that do not build upon each other.	1.0	1.0
The exercises were too difficult.	2.0	1.0
Through the mini lectures and practical exercises I feel well prepared for the implementation of the group project.	4.0	3.0
I liked that the exercises built upon each other.	4.0	4.0
The exercises were too big.	1.0	2.0

venue and are, therefore, reported here as a whole. Note that the following summary of students' answers to those questions is not a formal content analysis, but an identification of trends done by the author of this thesis. In the following, only trends mentioned by at least four students are reported on.

To the question *What I liked most about the plenum?*, 44 students provided an answer. Of those students, the vast majority (27 mentions) liked the practical exercises or the practical part of the lecture sessions most, expressed through statements such as:

- "The mini exercises during the lecture"
- "The practical exercises were an amazing opportunity to understand the topics"
- "Through the exercises one was able to understand nearly all of the theory"[2]

Furthermore, students positively mentioned the teams and the discussion within the teams (7 mentions), as well as the help of the lecturer during the exercises (7 mentions). 4 students spoke positively about the content of the course regarding difficulty and the chosen topics.

38 students gave an answer to the question *What could be done better in the future?*. Students mostly cited content issues (12 mentions) where students expressed the desire for more detailed explanations or to cover other topics. Problems with the JavaScript editor were mentioned as well (5 mentions), with four students explicitly stating the request to use a real development environment instead of the web-based JavaScript editor. Regarding the exercise design, students would have liked to have more time for the exercises or less extensive exercises (4 mentions). Additionally,

[2]Translated from German: "durch die Aufgaben konnte man wirklich fast die ganze Theorie verstehen"

students requested various improvements and features for Backstage 2 as their answers to this question (7 mentions).

Answers to *Other comments* were given by 11 students, and only a single theme (with an exception, see below) was mentioned three times, which was praise for the course, expressed through statements such as "it's the best practical exercise anyone could apply to :)" or "(q)uite motivating course!"[3].

Four students mentioned organizational matters, such as the grading of the final projects or better information about the course of the practical, but are disregarded here, as they do not pertain to the format.

Lecturer's Observations While supporting teams during the exercise phases, the lecturer, for one thing, used the class overview in **PCI2**, and for another, made a few observations on how teams used the JavaScript editor and their general approach to solving exercises. The latter observations only pertain to those teams which actually were supported personally by the lecturer during lecture sessions and are, therefore, not representative of the whole audience.

Some teams did not use the units that displayed the scaffolded exercises introduced in **PCI2** in the intended way. Those teams clicked the button that reveals the next subtask until all subtasks were displayed and started to work from there, which is an approach that completely voids the goal of supporting teams to focus on one subtask at a time.

Regarding the JavaScript editor, it was observed that teams did not use the tests at all or were unable to understand the error messages returned by the testing framework. Furthermore, the run- and compile time errors returned by the editor posed a problem as well, with teams often failing to even find the location in the code the error stemmed from.

In **PCI2**, the classroom overview was utilized to identify struggling teams for those exercises for which unit tests existed. It became quickly evident, that the number of passing unit tests over time or the slope of that graph are no appropriate measures for identifying struggling teams: A syntax error leads to the code failing all unit tests, and that, combined with teams generally requiring more than one run before fixing a syntax error, led to flatlining graphs and negative slopes even for teams that had already solved the majority of the exercise. Furthermore, the effect on the average slope is more pronounced when dropping from a high number of passing unit tests

[3]Translated from German: "Ganz motivierender Kurs!"

to zero as when dropping from a low number of unit tests to zero, that is, syntax errors had a more severe effect on the slopes of more successful teams.

The next section brings all the results reported above together, discusses the results and implications for Phased Classroom Instruction going forward.

Discussion

At the beginning of this section, three factors were mentioned as being important for the success of a format such as Phased Classroom Instruction: Students being able to solve the exercises, students requiring similar amounts of time to finish the exercises, and students liking the format. The following section discusses the results of the evaluations of **PCI1** and **PCI2** under these considerations.

Looking at the ability to solve the exercises correctly during the lecture sessions, students had problems across both venues with the majority of teams not even being able to solve correctly half of the exercises during the lecture sessions. The other aspect of exercises, the working time, showed in median a five minute difference between teams who were able to solve the exercise. While five minutes do not seem much on paper, that are five minutes in which still working teams and the lecturer are potentially disturbed by those teams. Scaling down the difficulty of exercises would most likely have a positive effect on correctness, but would change nothing regarding the differences in working time as more experienced teams would still finish earlier. Indeed, there seems to be a field of tension between these aspects, where optimization of one aspect leads to adverse effects on the other aspect. Hence, one aspect should be chosen for optimization, and ways to mitigate the negative effects of the other aspect should be conceived.

Improving upon the correctness seems more important than minimizing the differences between the working times, as being able to correctly solve the exercises promotes a sense of achievement among the students and is – especially in the case of the evaluated courses – important as exercises build upon each other. With regards to the exercises used in the courses, making those easier is hardly possible, as those were already reduced to their most important aspects when the duration of lecture sessions was cut down by 45 minutes. That leaves two areas for improvements: Splitting exercises down into smaller exercises and improving upon the technological support to enable more teams to correctly solve the exercises with just the support provided by the editor. Adapting the difficulty of the exercises (see Section 4.2.2) to a team is another approach but would be associated with a high effort, and hence, other means should be explored first.

While the results on correctness and working time are not especially convincing, the students' attitude towards Phased Classroom Instruction is all the more convincing across both venues: Students liked the format and its components (i.e., the exercises and the discussions within the teams) very much and vastly preferred Phased Classroom Instruction to a traditional lecture. The students' positive attitudes were further emphasized through their answers to the free text questions where more than half of the students explicitly mentioned the active parts and the exercises as the best part of the course. Moreover, the constantly high student attendance in the lecture sessions makes a case for the format as well: At the author's institute, attendance of that magnitude and consistency is not usual which suggests that students see the purpose in attending the lecture sessions.

Coming back to the students' attitudes, students found the course material to be sufficient to solve the exercises and found the exercises neither too difficult nor too extensive and liked that the exercises built upon each other. Touching upon exercises building upon each other, there was no difference in students' attitudes between **PCI1** and **PCI2** in statements referring to that property of the exercises, even though the scaled-down exercises in **PCI2** did only result in the shell of a game and not a playable game, such as Snake. That suggests that not the goal but the process of working on an ever-evolving piece of code is what was liked by students.

Students found Backstage 2 to support the format well, but issues with the JavaScript editor were revealed through the evaluations: While students found that the editor helped them getting started with JavaScript and was easy to operate, the majority of students would have preferred to use a real development environment for solving the exercises. That negative aspect is further reinforced by students mentioning the JavaScript editor negatively in their answers to the free text questions. As web-based editors enable the class overview and provide scaffolding and immediate feedback to students, they are an integral part of Phased Classroom Instruction which eliminates the possibility of allowing students to work on the exercises using a development environment of their choice. Therefore, the editor itself has to be improved to better its acceptance among students.

Which parts of the editor to improve is unfortunately not evident from the surveys, as the surveys did not include questions dedicated to that, but a few improvements can be derived from the lecturer's experience: Error messages, both those returned from the JavaScript interpreter as well as those returned from the unit testing framework, have to be made more understandable. Moreover, the interface for scaffolded exercises was not used in the intended way, which leaves room for improvement in that area as well.

As the lecturer felt that the number of passing tests and the slope of that graph were not able to predict which teams require help, the class overview has to be reworked. Removing the time component and simply looking at the unit tests that were passed might be a better approach.

This section closes the evaluations of the first two evaluations of Phased Classroom Instruction which revealed much room for improvement, mostly related to the technological support. The next section first discusses the changes made to the course material and the technological support and then presents the results of an evaluation of Phased Classroom Instruction in a further venue of the software development practical.

6.4 Going Further with Phased Classroom Instruction

This section first introduces the adaptions that were made in response to the results of the previous evaluations and then presents the results of a third evaluation.

6.4.1 Adaptions to the Technological Support and Course Material

While adaptions to the exercises were made in response to the previous evaluations, the main adaptions were made to the technological support. The scaffolded exercises were integrated into the editor and the error reporting of the editor was improved. Exercises were reworked and tests were constructed in a scaffolded way for all of the exercises.

Updated JavaScript Editor

The following section describes an updated version of the JavaScript editor, which was developed by Anna Maier [Mai19] as part of her master thesis. Furthermore, this section shortly outlines an evaluation of the previous terms' submissions to identify common errors made by students which was done by Maier as well. Refer to her master thesis for a more detailed description of the editor and the study.

Improvements to the JavaScript editor mostly focussed on two areas: The subtasks were integrated directly into the editor so that students can leverage the scaffolding

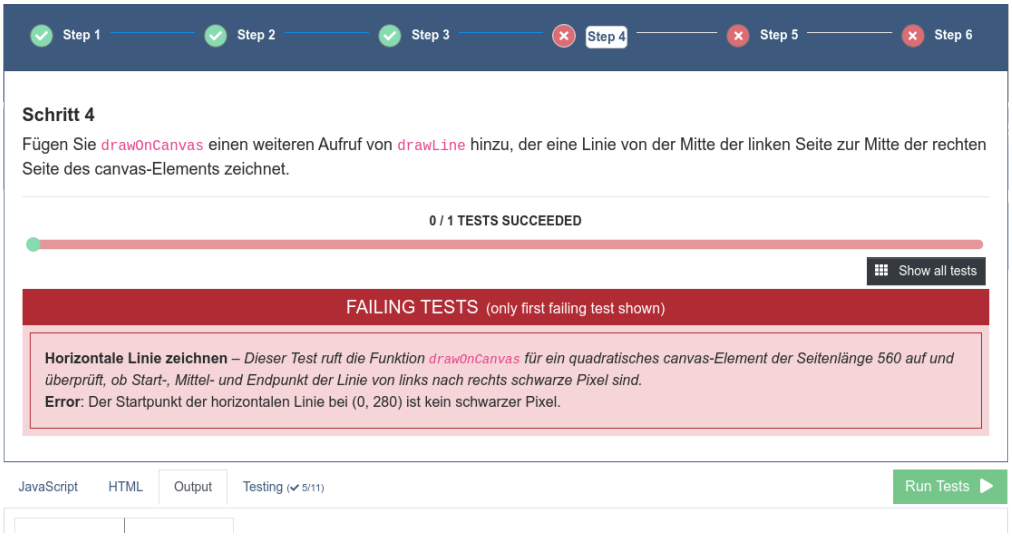


Fig. 6.9.: A screenshot of the scaffolding provided by the updated JavaScript editor: At the top, the integrated subtask interface can be seen and below that the current step and its progress is shown. At the bottom the JavaScript editor can be seen.

provided by them more easily, and the error reporting was improved to both make it easier to understand the error as well as to identify errors before running the code.

A screenshot of the integration of subtasks into the editor can be seen in Figure 6.9. Each step at the top represents one subtask, a green checkmark or a red cross next to each of them represents whether all tests associated with that subtasks are passing or not. The currently worked on step is highlighted with a white background and can be changed at any time regardless of tests in previous steps failing. Beneath the overview of all steps, the description and the progress (expressed as the percentage of passing unit tests) of the current subtask is shown. Only the first failing test associated with the current subtask is shown; an overview of all tests of that subtasks is only shown after clicking on the button labeled *Show all tests*. Below that, a part of the JavaScript editor introduced in Section 6.3.1 can be seen.

While the part of the editor where code is entered remained visually mainly unchanged, various changes were made to the inner workings of that part: The first version only provided basic error messages on syntax errors while typing, and hence, these were the only errors students could catch before running the code. That was extended in the updated version to include error messages for common JavaScript mistakes as well. For identifying those mistakes, ESLint,[4] a static code analysis tool for JavaScript, was used. Among the errors ESLint can identify before running the code are using identifiers before their declarations[5] and using = (assignment) instead of == or === (comparison) in the conditions of if-clauses and loops.[6] As

[4] `https://eslint.org/`
[5] `https://eslint.org/docs/rules/no-use-before-define`
[6] `https://eslint.org/docs/rules/no-cond-assign`

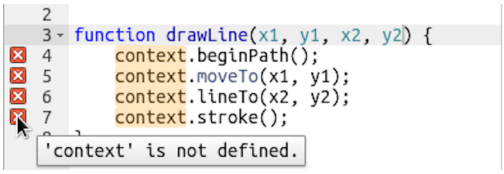


Fig. 6.10.: Example for an ESLint error message shown in the editor: The red rectangle and the yellow marking on the code identify the part of the code were the error was found. Hovering over the rectangle reveals the error message.

Fig. 6.11.: Reporting of run- and compile time errors in the updated JavaScript editor: After clicking on a line in the stack trace, the location of the error is highlighted in the text area below.

ESLint can identify far more errors than are relevant for programming beginners, the submissions of the previous venues were used to identify the errors students made in those venues. The most common errors from that evaluation combined with errors found in other research on common errors made by programming beginners are in the updated version of the editor immediately highlighted while users enter their code. An example for an ESLint error reported by the editor before running the code can be seen in Figure 6.10: In the example, the used identifier `context` is not declared, something that in the previous version would have only produced an error when running the code.

Turning towards errors returned by JavaScript itself, the interface was reworked to make the connection between the stack trace and the code more evident. Similar to a real development environment, clicking on a part of the stack trace brings the respective line into focus and highlights the part of that line the error resulted from. An example for that can be seen in Figure 6.11: In the example, a function `setupKeyboardListener` that is not defined for the class `Game` is called what results in a runtime error. The red highlight in the text area reveals the exact location of the error.

Updated Class Overview

As identified in the evaluation of **PCI2**, the class overview turned out to be not suitable for determining struggling teams. The following section introduces the updated version of the classroom overview, which was also developed by Anna Maier [Mai19] as part of her master thesis.

In the updated version of the class overview, teams are ordered dynamically based on their predicted need for a lecturer's personal support. For determining the team order, four attributes are used:

1. *Exercise already solved*, is true if there is one previous submission in which all unit tests were passed; false otherwise
2. *Percentage of passing tests*, the percentage of unit tests the current submission is passing
3. *Current step*, the current exercise step selected by the team
4. *Unsuccessful compiles*, the number of unsuccessful (i.e., leading to a compile time error) runs since the last successful (i.e., not leading to a compile time error) run

The attributes are used in the order they appear in the list above when determining the position of a team in the class overview: First, teams that have already solved the exercise are sorted to the bottom, then teams inside each of those groups (i.e., teams having solved the exercise and teams not having solved the exercise) are sorted by their percentage of passing tests with teams with a lower percentage being sorted to the top, and so on with the other attributes. Hence, the position of an attribute in the list above represents the importance that attribute has on the order of teams in the overview.

The reworked overview can be seen in Figure 6.12: Each team is represented by a row of the table, and each cell of a row (except for the first and last) represents a test. A check indicates that this test was passed the last time that code was run (in the following simply called run), a cross that this test failed the last run. The number next to each failing test represents for how many runs that team failed that test. Thus, the updated overview removes the visualization of the time component and focusses on a team's current run. The current step selected by a team is shown by displaying the results from tests associated with that step in a larger font (e.g., Team AA is currently working on step 4). Furthermore, the overview contains information about the average time required for each step (the number next to each step in the

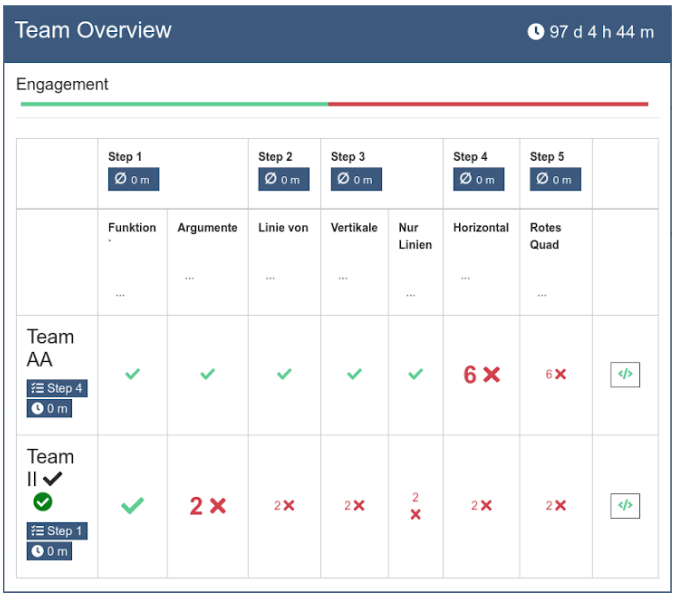


Fig. 6.12.: Screenshot of the updated version of the class overview: Each team is represented by a row, which shows the current step (larger font size) and the passing and failing tests (check and cross, respectively). Furthermore, the average working time per step (the number next to each step), as well as the current working time of the team in the current step (the number next to each team), is shown.

top row) and for each team, the time they are working on the current step (number below the team name).

The course material, especially the exercises and the associated unit tests, were reworked for the third venue to leverage the new possibilities provided by the updated JavaScript editor and class overview.

Updated Course Material

To leverage the possibilities of the new editor, two changes were made to the exercises: While the goal of the exercises remained nearly unchanged compared to **PCI2**, the descriptions of the individual subtasks were made less abstract to provide more guidance (e.g., "Implement a method `changeDirection` that is called when a key is pressed" to "Implement a method `changeDirection` and register that function as an event listener for the event `keyup` in the constructor"). One exercise, *3-2*, contained two related concepts (updating the game in regular intervals and variable movement speeds of game objects) but was solved by few teams successfully in **PCI2**. Hence, that exercise was split into two exercises. Furthermore, updating the game in regular intervals is now implemented using `window.setInterval` instead of `window.requestAnimationFrame`, which might be a less optimal solution but is much easier to implement.

More extensive changes were being made to the units tests of the exercises: Tests were rewritten to provide error messages in natural language instead of the error returned by the testing framework, and tests were ordered in a scaffolded way, so that students could follow a trail of failing tests, passing test by test, in order to arrive at a correct submission.

Starting with the error messages returned by the unit testing framework, which, from the lecturer's experience, have not been understood well by students. Indeed, the error messages returned by the testing framework and shown to students were in many cases rather arcane. An example of an error message can be seen in Figure 6.4 at the beginning of the chapter. While the title and description explained for each test what was being tested, the error message did not clearly indicate what went wrong. Taking the error message of the failed test in the figure as an example, the error message "expected 0 to be above 0" has seemingly nothing in common with the task of drawing a line. In reality, that unit test checked for a point on a line that should have been drawn by the user's code (that is, having RBG values over 0), if that point was non-white (by checking whether the RBG values of that point are above 0). The test failed because the point was still white, that is, its RBG values were 0. Hence, if used at all, those tests could have served as indicators for a complete and correct submission, but not to guide students towards a correct submission. For that reason, all tests were rewritten to provide error messages in natural language and, if possible, cues on how to proceed. Those error messages in natural language ranged from simple messages, such as "The call of `hello(x)` returned `Y` and not the expected value `X`", to more complex messages, such as "The function added as event listener is not executed in the correct context. Did you use `bind` or an arrow function?".

Coming to *scaffolded tests*, tests were written for each of the subtasks so that students were able to notice when a subtask was solved correctly and ordered to provide an additional layer of scaffolding: Tests build on each other, that is, later tests require previous tests to succeed. For example, a first test asked for the implementation of a function `foo` and the consecutive test to implement (part of) the functionality of `foo`. Furthermore, tests were ordered in the same way the subtask description asked for components to be implemented, so if the description of a subtask asked for the implementation of a function `foo` and then for the implementation of a function `bar`, the test for `foo` came before the test for `bar`. Scaffolded tests leverage two features of the updated JavaScript editor: First, the association of unit tests with subtasks and that, by default, only the first failing test is displayed. In that way, students can always focus on the currently failing unit tests and go from failing test to failing test without being distracted by a long list of failing tests.

With all those improvements to the technological support and the course material, the course was run again in the winter term 2019/20 with 60 students. The following section presents the evaluation of that venue.

6.4.2 Study

The third venue, **PCI3**, was nearly identical to **PCI2** but used the improvements to the technological support and course material described above. Furthermore, seven instead of six lecture sessions were held, as due to the greater number of students more time was required for organizational matters during the first lecture session (such as assigning groups, finding dates for weekly meetings, ...).

Methods

PCI3 was evaluated using the same methods as described in Section 6.3.2 with the following additions to the survey. The entire survey can be found in Appendix A.2.

- A block of seven questions measuring the students' attitude towards the JavaScript editor and the scaffolded tests to be answered using the same six-point Likert scale as the other exercises. Five of those questions were adapted from the Maier's [Mai19] survey.
- Two additional questions to the block measuring the students' attitude towards the course material asking if the exercises were too easy and if the lecturer was always there when help was required during the exercise phases.

The correctness of the exercises was determined the same way as in **PCI2**.

When comparing more than two samples, the Kruskal-Wallis H-test with a significance threshold of $p = 0.05$ was used, as the data does not follow a normal distribution which calls for a non-parametric test (see [CF14]). Post-hoc testing was done using the Mann-Whitney U test (see [CF14]) with Bonferroni correction (see [EW07]). In this part, in case of post-hoc testing, three comparisons take place (between **PCI1**, **PCI2**, and **PCI3**), and hence, the significance threshold was adapted using Bonferroni correction to 0.017 (dividing the regular significance threshold by the number of comparisons). Aggregated measures are reported as Median, hereafter abbreviated as Mdn, as it is more robust against outliers [How09], and therefore, deviation is reported as Median Absolute Deviation, hereafter abbreviated as MAD (see [RC93]).

Tab. 6.8.: Overview of the population of **PCI3**.

Course	# of students	# of survey participants	avg. coding proficiency
PCI3	60	41	3.7

Results

Table 6.8 shows an overview of the population of **PCI3**. Similar to the previous venues, the majority of students took part in the survey, but as the survey was conducted during the last section, again, only the opinions of students being present in that session could be collected. The average coding proficiency between the three venues is not significantly different as indicated by the Kruskal-Wallis test ($p = 0.84$).

Exercise Correctness and Working Time Looking at the results on how successful teams were in solving exercises correctly during lecture sessions, shown in Figure 6.13, teams were much more successful in solving the exercises compared to the previous two venues (Mdn: 83.3%, MAD: 12.4%). The Mann-Whitney U test indicates that this increase is significant compared to **PCI2** ($p = 0.0003$). Comparison with **PCI1** is omitted due to the huge differences in the exercises.

With the exception of one team, all teams in **PCI3** solved at least half of the exercises correctly during the lecture sessions. Generally, exercises are solved correctly by at least 60% of all teams with the exercise on object-oriented programming, *5-1*, being an exception again. Nonetheless, teams did much better in that exercise compared to **PCI1** and **PCI2** with 4 teams solving the exercise correctly and 7 teams being very close to a correct submission with only one or two small details missing.

The working time until the first correct submission can be seen in Figure 6.14 and still varies greatly across teams and exercises. Contrary to the figure for **PCI2**, this figure can be taken at face value, because the majority of teams solved the exercises correctly. There are still exercises in which the differences between the teams are small, such as *2-1* and *6-2*, but generally, the difference is rather large. The interquartile range is comparable to **PCI2** with a median of 6.3 minutes and ranges from 2.7 to 13.1 minutes. Standing out in this figure is the existence of a no small number of outliers which were not present in the figure for **PCI2**.

In summary, the exercise correctness was significantly improved compared to the first venues, but working time still varied greatly with the presence of outliers not being present in the working times of **PCI2**.

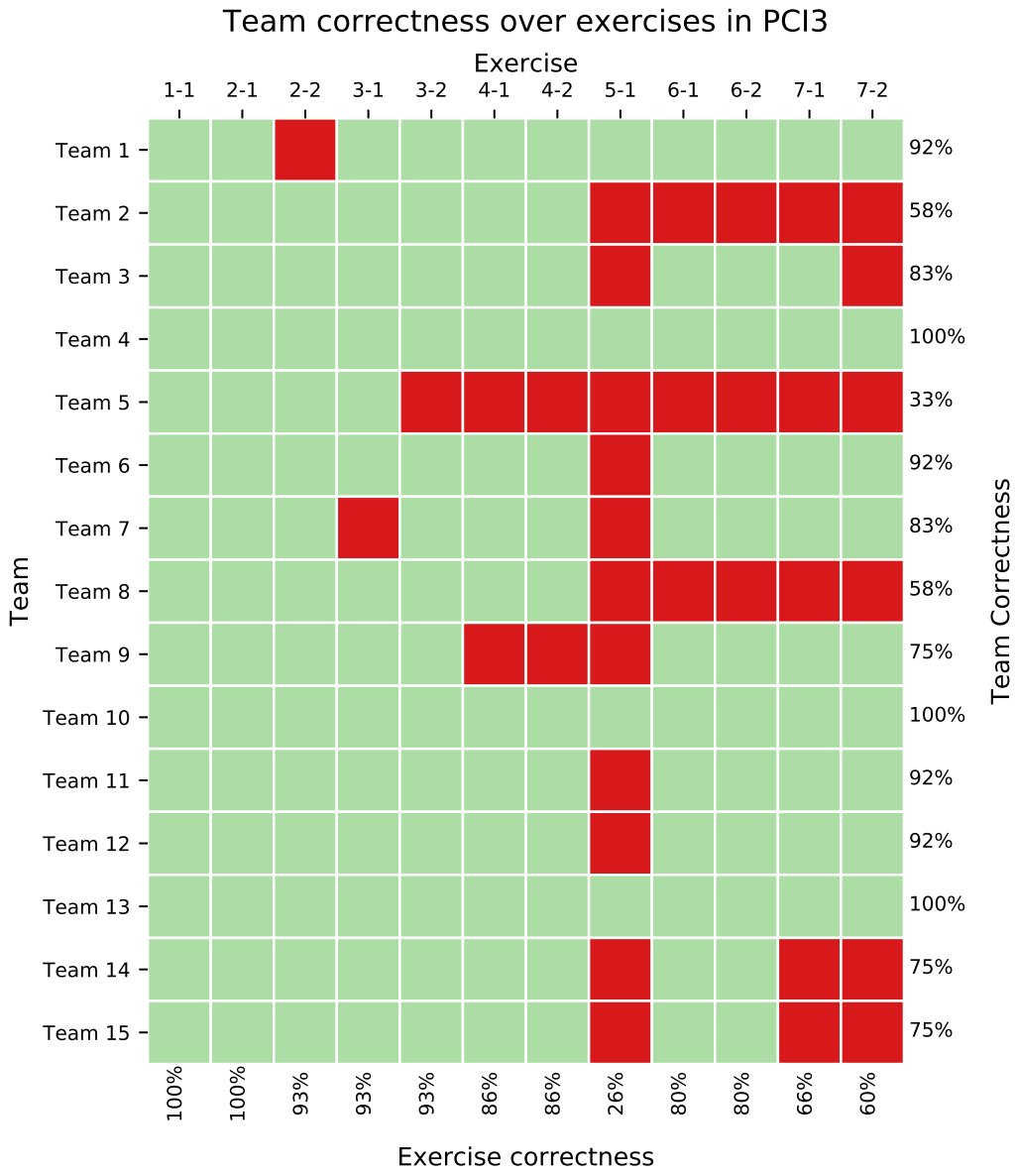


Fig. 6.13.: Overview of exercises being solved correctly during lecture sessions by team and exercise for **PCI2**. A green square indicates that the team was able to solve the exercise correctly during the lecture session; a red square that the team was not able to solve the exercise during the lecture session.

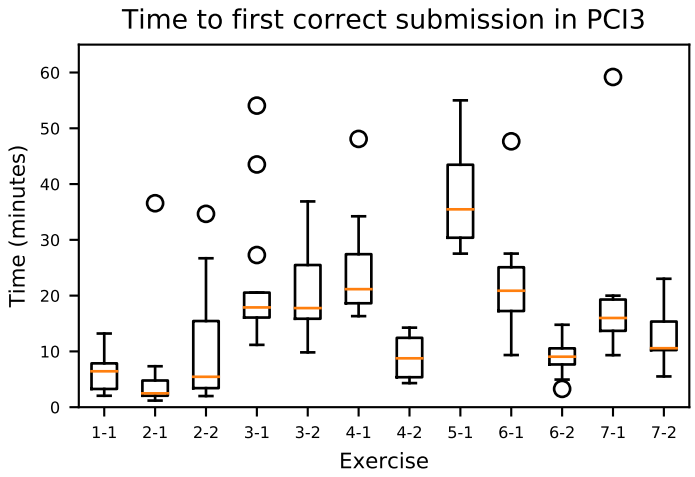


Fig. 6.14.: Time of first correct submission for each exercise in **PCI3**.

Tab. 6.9.: Percentage of students being present during lecture sessions as counted by the lecturer in **PCI3**.

Course	L1	L2	L3	L4	L5	L6	L7
PCI3	95%	95%	77%	77%	72%	82%	70%

Attendance during Lecture Sessions In Table 6.9, the percentage of students being present during the lecture sessions as counted by the lecturer can be seen. Note that the attendance for the first lecture was not counted, but from the lecturer's memory, there were at most 3 students not present. Attendance started high, similar to **PCI2**, and then began to drop, but evened out at attendance rates of between 70% and 80%.

Students' Attitude towards Phased Classroom Instruction Looking at the results of the blocks measuring the students' attitudes towards Phased Classroom Instruction and the course material, which can be seen in Tables 6.10 and 6.11, respectively, confirms the findings made of the previous two venues: Students liked the format, its components, and the course material. Students found the exercises not too extensive and appropriate regarding their difficulty, as they found the exercises neither too difficult nor too easy. The majority of students (25 of 41) strongly agreed with the statement that the lecturer was always there when the team had problems solving the exercise.

Turning towards the students' attitudes towards the technological support, which are reported in Table 6.12, shows that the results for this block are nearly identical to the previous venues though there is one exception: Contrary to the students in **PCI1** and **PCI2**, students in **PCI3** rather disagreed with the statement that they would have liked to work on the exercises using a real development environment. The

Tab. 6.10.: Results of the survey block measuring the students' attitude towards Phased Classroom Instruction for **PCI3**.

Statement	PCI3 Mdn
The immediate practical exercises after the mini lectures helped me understand the topic.	5.0
Discussions with my team mates during the practical exercises helped me understand the topic.	5.0
I would have preferred a traditional lecture without practical exercises.	0.0
I had fun during the plenum sessions.	4.0

Tab. 6.11.: Results of the survey block measuring the students' attitude towards the course material for **PCI3**.

Statement	PCI3 Mdn
The mini lectures were sufficient to solve the practical exercises.	4.0
I would have preferred exercises that do not build upon each other.	1.0
The exercises were too difficult.	1.0
Through the mini lectures and practical exercises I feel well prepared for the implementation of the group project.	3.0
I liked that the exercises built upon each other.	4.0
The exercises were too big.	1.0
The exercises were too easy.	2.0
The lecturer was always there when my team had problems solving the exercise.	5.0

Tab. 6.12.: Results of the survey block measuring the students' attitude towards Backstage 2 for **PCI3**.

Statement	PCI3 Mdn
The JavaScript editor on Backstage made the getting started with JavaScript easy.	4.0
The JavaScript editor was easy to operate.	4.0
The interface of Backstage, where exercises were worked on, was clearly designed.	4.0
The course format (i.e., mini lectures, followed by exercises and peer review) was well-supported by Backstage.	4.0
I would have preferred to solve the practical exercises using a real development environment.	2.0

Tab. 6.13.: Results of the survey block measuring the students' attitude towards the updated editor and exercise design for **PCI3**.

Statement	PCI3 Mdn
The yellow markings on the code and the accompanying error messages helped to identify error before running the code.	4.0
It was clear to me at what point my team should move on to the next exercise step.	4.0
The error messages in natural language helped to recognize errors in my group's code.	4.0
I would have preferred to see all failing tests instead of only one failing test.	2.0
Switching between exercise steps was easy.	4.0
Even without the error messages in natural language, I would have been similar fast in recognizing errors in my group's code.	2.0
Seeing only one exercise step helped focus solving that exercise step.	4.0

Kruskal-Wallis test indicates a significant difference between the venues in respect to that statement ($p = 0.016$). Post-hoc testing using the Mann-Whitney U test (remember, α adapted to 0.017 using the Bonferroni correction) reveals a significant difference between **PCI2** and **PCI3** ($p = 0.005$) with the comparison between **PCI1** and **PCI2** being just a bit above the adapted significance threshold ($p = 0.019$).

The results of the fourth block, which measured the students' attitudes towards components of the new editor and the scaffolded tests, can be seen in Table 6.13. Students found the yellow markings (i.e., the error messages generated by ESLint) helpful for identifying errors before running their code, liked to see only the first failing test, and liked the error messages to tests in natural language. Moreover, seeing only a single step helped students to focus on the task at hand, and the interface elements accompanying the steps were clearly designed.

Overall, the students' attitudes in **PCI3** further confirm the results of the previous venues, as they are nearly identical to the results of those venues. A single difference exists concerning the JavaScript editor with students in **PCI3** preferring the web-based JavaScript editor over a real development environment for solving the exercises. Additionally, the components of the updated editor and the scaffolded tests with error messages in natural language were well-liked by students.

Students' Opinions The survey conducted in **PCI3** contained the same free text questions as the surveys conducted in **PCI1** and **PCI2**. The students' answers to these free text questions were evaluated in the same way as described in Paragraph 6.3.2.

To the question, *What was liked most about the lecture sessions?*, 36 students provided an answer, and similarly to **PCI1** and **PCI2**, the majority of students mentioned the practical exercises expressed through statements such as "I liked to immediately apply the concepts. I learned faster that way"[7]. Related to that, students made positive statements regarding the course design through statements such as "fast shift between theory and practice"[8] or "short theory – much practice"[9].

While only mentioned in two statements, students positively mentioned the JavaScript editor which was not mentioned positively at all in the students' answers in the previous venues.

26 students answered the question *What could be done better in the future?*. In their answers, they mostly requested features for Backstage 2 (10 mentions), with five of those students requesting the same feature: Automatic copying of the previous exercise's code into the editor for the current exercise. Indeed, students were required to first navigate to the previous exercise, copy their code, navigate to the current exercise, and paste their code before they could start working on the current exercise. Besides that, students mentioned content issues (8 mentions), such as errors on the slides or requested to cover additional topics or treat some topics in more detail.

As *Other comments*, five students provided a statement. Only one theme was mentioned more than once which were positive comments about the course (3 mentions), such as "all those tests: great work!"[10] or "very good format! (exercises in the lecture sessions)".[11]

[7] Translated from German: Ich fand es gut, die Konzepte direkt praktisch umzusetzen. Dadurch habe ich schneller gelernt
[8] Translated from German: schneller Wechsel zwischen Theorie und Praxis
[9] Translated from German: kurze Theorie - viel Praxis
[10] translated from German: Die ganzen Test: Tolle Arbeit!
[11] Translated from German: sehr gutes Format! (der Übungen in der Vorlesung)

Lecturer's Observations In the lecturer's experience, the updated classroom overview made it easier to identify which teams required help. A typical exercise phase was conducted as follows: After starting the exercise, the lecturer waited for a few minutes and then began to visit the team currently at the top of the class overview. In the majority of cases, the team being visited did not decline the lecturer's support and had a problem. In between, the lecturer supported teams asking for help on their own initiative. Indeed, teams were much more inclined to ask for help on their own initiative in this venue compared to the previous ones.

While looking through all teams' submissions to identify the first correct submission of each team, it stood out that a few teams exhibited a *tests-only* approach to solving the exercises. Such teams considered an exercise as complete as soon as all tests were passing even though their code had no output at all or output that contained clearly visible errors. Examples for such behavior are never instantiating the main class which results in no output at all (3 teams in exercise *5-1*) or not clearing the canvas which results in all game objects leaving a trace behind (2 teams in *5-1*). The former was not covered by tests, but mentioned in the respective subtask description, and the latter was tested only superficially.

Besides that, it was noticed that one or two teams started to explore alternative approaches to solving an exercise after having already solved the exercise correctly, such as implementing a functional approach after first solving the exercise using an imperative approach. That explorative behavior led to a few false positives on teams requiring help during the first lecture sessions as those teams were shown at the top of the list as their codes were passing no tests even though they had already completed the exercise. That behavior of the system was remedied by the third lecture sessions with teams already having solved the exercise being automatically sorted to the bottom of the class overview (see the first order criteria described in Section 6.4.1).

Discussion

The following section discusses the results of the evaluation of Phased Classroom Instruction in **PCI3** and discusses especially the aspects exercise correctness during lecture sessions, differences in the working time until first correct submission between teams, and attitudes of students towards Phased Classroom Instruction.

First things first, even with 60 students, the results of **PCI1** and **PCI2** regarding the students' attitude towards Phased Classroom Instruction could be reproduced. Students liked the format, its components, and vastly preferred Phased Classroom

Instruction to traditional lectures. This is, again, further reflected by the high attendance to the lecture sessions which, same as in the first two venues, never dropped below 70% in **PCI3**.

The number of exercises solved correctly during the lecture sessions increased significantly from around 44% in **PCI2** to around 83% in **PCI3**. As mainly two aspects were changed from **PCI2** to **PCI3**, it is unclear to which parts that increase can be attributed to the updated JavaScript editor and the scaffolded tests. Taking the positive attitude of students towards both aspects into account suggests that both changes played their part: Regarding the editor, students found that the step interface helped them to focus on the current step and that the error messages provided by the editor helped them to find errors before running their code. Furthermore, students in **PCI3** significantly less preferred to use a real development environment as opposed to the JavaScript editor than students in **PCI2** which suggests that the updated editor provided students additional value which was not provided by the previous version. Regarding the scaffolded tests, students found that the error messages in natural language helped them to solve the exercises faster.

It would be detrimental to students' learning if the exercises would have gotten too easy through the changes made for **PCI3**, but as students felt that the exercises were neither too difficult nor too easy that concern can most likely be dismissed. Nonetheless, there were teams for which the exercises were easier and teams for which the exercises were more difficult as indicated by the varying working times until the first correct submission.

Regarding the varying working times on exercises, it is most likely inevitable that more experienced teams generally finish exercises earlier than less experienced teams. Hence, ways to engage those more experienced teams beyond the completion of an exercise have to be found so that those have something to do and do not disturb the classroom. One possibility to engage those teams is getting them to help other teams, a form of peer teaching, which could, for example, be done by showing suggestions which teams to help after a team's code passes all unit tests. Another means for engagement is to encourage the aforementioned explorative behavior that was noticed by the lecturer which could be promoted by adding open subtasks at the end of an exercise.

A last point regarding the working times is the presence of outliers in **PCI3** which were not present in the working times of **PCI2**. With one exception, those outliers are teams that took considerably longer than other teams to complete the exercise but completed the exercise nonetheless. The absence of such outliers in **PCI2** suggests that those teams would not have been able to solve the exercise without the improved technological support in **PCI2**, as those teams most likely simply would not have

finished the exercise in **PCI2**. Hence, the updated technological support succeeds in empowering more students to successfully solve the exercises during lecture sessions.

The updated class overview turned out the be more useful for identifying whom to help in the lecturer's experience. While the updated overview still had a time component, as it always represented the current run of a team's code, the reduced view without history on previous runs (in form of graphs) and the dynamic ordering of teams might be the factors that improved its usefulness.

Summarizing the results from the evaluation of **PCI3**, the adaptions made to the technological support and the course material turned out to be a success: Students were significantly more successful in solving the exercises during the lecture sessions compared to **PCI2**, and their positive attitudes towards Phased Classroom Instruction remained unchanged. Both the editor and the scaffolded tests did their part in improving the students' success. Results further point towards the importance of integrating subtasks in the environment the exercise is being worked on and providing a concrete goal to work towards. Furthermore, providing error messages in natural language supports students in solving coding exercises more successfully.

This discussion closes the section on the evaluation of Phased Classroom Instruction in the third venue of the software development practical on JavaScript programming. The final section of this chapter brings all results together, discusses implications for Phased Classroom Instruction and future research avenues.

6.5 Wrapping up Phased Classroom Instruction

In this final section, the findings for Phased Classroom Instruction are summarized and future research avenues and perspectives are presented. Phased Classroom Instruction is a learning and teaching format that combines mini-lectures with extensive exercises in which students immediately apply the just taught knowledge, optionally followed by peer review. Phased Classroom Instruction addresses various issues of a similar learning and teaching format, flipped classrooms: Among the issues of flipped classrooms is the high time effort associated with the production of the learning material (mostly in form of videos) using which students acquire the knowledge outside lecture sessions and that flipped classrooms scale badly to larger course sizes as there are only so many students which can be supported effectively by a lecturer. Mini-lectures in Phased Classroom Instruction address the effort as those are held by the lecturer and do not have to be produced as opposed to videos, and students and lecturers are being supported by technology to make the format

scale: Students are working on the exercises using problem- or subject-specific editor that provide immediate feedback and scaffolding, and lecturers are supported by an overview of the class to help them identify struggling students.

Among the requirements for Phased Classroom Instruction is that students are able to solve the exercises, finish the exercises temporally close to each other, and finally, that they like the format. This chapter presented three venues and the adaptions and improvements made between them to achieve the objectives of the format. As the evaluations of Phased Classroom Instruction done as part of this work focussed exclusively on computer science education, many of the findings are specific to that area but some of them can be generalized for applications of the format in various STEM subjects as well.

Overall, Phased Classroom Instruction worked well to bring active learning even to large classes – throughout the three evaluated venues of the same software development practical, the number of participants was scaled up from 16, to 44, to finally 60. Across all venues, the students' attitudes towards the format remained unchanged with them liking the format and its components very much and vastly preferring the format to traditional lectures. While in the first two venues, students were rather unsuccessful in solving the exercises correctly during the lecture sessions, technological adaptions made for the third venue led to a significant increase in teams' performances. The working times until the first correct submission still varied greatly between the different teams but suggested that the adaptions for the third venue empowered less experienced teams to successfully solve the exercises during the lecture sessions. A fourth venue was held during summer term 2020 with 84 participants but was due to time constraints not formally evaluated. However, that venue is shortly reported on in Chapter 10.

Two major changes were made for the third venue to which the significant increase in students' performances most likely can be attributed to: First, the subtasks of the exercises were integrated directly into the editor and common programming errors were displayed immediately in students' code which added a new layer of immediate feedback. Second, tests accompanying each exercise were ordered in a scaffolded way and written to return error messages in natural language explaining what went wrong and giving cues on how to proceed. Those tests were ordered to conform to the implementation order suggested in the respective subtask as well as written to build upon each other. Those two choices leveraged that the interface of the editor by default only displayed the first failing test – allowing students to solve the exercise by just going from failing test to failing test. These changes are not limited to programming in JavaScript but can be transferred nearly literally to other venues of programming education as well.

While the changes made for the third venue addressed the problem of teams' performances, the teams' working times still varied strongly. Hence, to prevent already finished teams from getting bored or disturbing the classroom, two activities for those teams were suggested: Get them to help other teams or nudge them to further experiment with their code.

Leaving implications for computer science education and coming to Phased Classroom Instruction in general, the format should be applicable in most STEM subjects as long as the covered topics allow for problem- or subject-specific editors. The results of Phased Classroom Instruction so far suggest that is is important to break down exercises into subtasks and provide clear goals for each subtask. Further promoting performance is immediate feedback on the current submission in natural language. For the class overview to work, measures that represent a team's success have to be available for the exercise. Whether measures independent from the exercise are sufficient support lecturers to identify struggling teams is a perspective for future research. In the same vein, the class overview itself has to be formally evaluated as the current results are building upon a single lecturer's experience in a single course.

Evaluating the format in other contexts is a further research avenue: One of the most obvious contexts are tutorials (or lab sessions) that most often accompany a traditional lecture in which students are supposed to apply the concepts acquired during lecture sessions on the basis of exercises. At the author's institution, the ever-increasing number of students has led to those active learning opportunities slowly degrading to small lecture sessions where a tutor demonstrates how to solve the exercises on a blackboard. Phased Classroom Instruction can be used to make those tutorials active again with the lecturer demonstrating how to solve an exercise of that type (the *mini-lecture*) and students afterward solving another exercise of that type using a problem-specific editor.

As a last research perspective, peer review was not evaluated after the first venue due to time constraints (see [MB19c] for an evaluation of the peer review of **PCI1**). A more extensive evaluation of peer review has to be made to find evidence for the effectiveness and feasibility of an immediate peer review of extensive exercises during lecture sessions.

The greatest limitation of the presented evaluations is that they all took place in the same course and were taught by the same lecturer. Hence, evaluations of Phased Classroom Instruction in different contexts and with different lecturers are important to further validate the results. Furthermore, a few findings are based on a single lecturer's (anecdotal) experiences and have to be evaluated formally. As the survey was conducted during the last lecture session of the respective venue, only opinions

of those students present during that session were collected. It can be assumed that those students were generally more positive towards the format than students who were not present during that lecture session which might have introduced a positive bias to the results. Notwithstanding, as a great majority of students were present during the last lecture sessions, that bias most likely did not influence the results much. However, the opinions of students not present during the last lecture session might give valuable insight to improvements to Phased Classroom Instruction and should be collected (if possible) in future venues.

Phased Classroom Instruction is a format in which the lecturer goes "[f]rom [s]age on the [s]tage to [g]uide on the [s]ide" [Kin93, p. 30], and students play the leading part. In the next format, *Collaborative Peer Review*, lecturers are not even longer guides, but just facilitators for a format that is mainly student-centered.

7 Collaborative Peer Review

This chapter is based on the following article:

- Sebastian Mader and François Bry. "Towards an Annotation System for Collaborative Peer Review". In: *International Conference in Methodologies and Intelligent Systems for Techhnology Enhanced Learning*. Springer, 2019, pp. 1–10

In addition to the contents of the article, this chapter adds with **CPR4-10** seven further courses to the evaluation and extends the evaluation.

While in *Phased Classroom Instruction*, instructors still had the role of a "guide on the side" [Kin93, p. 30], in the following format, *Collaborative Peer Review*, the role of instructors becomes more like a "manager on the side" who simply orchestrates the format. As the name suggests, peer review is at the heart of Collaborative Peer Review, which is "an arrangement in which individuals consider the amount, level, value, worth, quality, or success of the products or outcomes of learning of peers of similar status" [Top98, p. 250]. As already discussed in Chapter 1, peer review can be utilized in face of large numbers of students to provide formative feedback in a timely manner and, more importantly, feedback at all.

Peer review of writings, such as essays, is often done in phases: First, the document to be reviewed is created which is then distributed to several reviewers. Each of the reviewers creates a review which are then returned to the author of the document (see, e.g., peerScholar [Col+15] and Mechanical TA [Wri+15] as examples for peer review systems with strict phases). A drawback of that approach is that authors are unable to inquire about the received reviews in case of ambiguities or disagreements. Furthermore, each reviewer reviewing on their own precludes any form of collaboration between them what potentially leads to the same work being done twice or more and potential disagreements between the reviewers remaining unresolved. Lastly, the document and its review are often restricted to authors and reviewers even though both are valuable resources for other participants as well, as they provide insights on how and in what quality others are creating their reviews and essays.

As discussed later in this chapter in detail (see Section 7.1), there are peer review systems where students can provide feedback on their received reviews and reviewers are expected to act upon that feedback, but only very few systems offer fully

bidirectional communication, and even then, only for a subset of the stakeholders of the review process. To the author's best knowledge, no peer review system enables bidirectional communication between all stakeholders during the review. Furthermore, only few systems allow all students to access other students' works and reviews (see, e.g., CritViz [Tin+13], SWoRD [CS07], and the system described by Hsia et al. [Hsi+16]).

Collaborative Peer Review addresses the aforementioned issues of traditional peer review of writings. In Collaborative Peer Review, all essays are provided as units in a course and for each essay, several students are tasked to create detailed reviews. Reviews are done as annotations using the collaborative annotation system (see Chapter 3) and are synchronized immediately among all other participants of the course, that is, the author, reviewers, and all other participants. Hence, in combination with the option to vote and comment on annotations, the immediately shared annotations allow for communication and collaboration between author and reviewers, reviewers, and participants unrelated to the essay as well. Note that Collaborative Peer Review is an explorative format – it was not planned as a format but was created after observing the interaction of students when Backstage 2 was used for peer review in three courses and to some extent in the course described in [MB18a].

In this chapter, the results of the evaluation of Collaborative Peer Review across ten seminar-style courses are presented and discussed. In seminar-style courses, every participant is assigned a topic, tasked to research that topic on their own, write an essay, and prepare a presentation. Finally, every student holds the prepared presentation about their topic in front of the other participants of the course.

This chapter is structured as follows: First, approaches to communication in peer review found in various peer review systems are discussed, followed by a more detailed introduction of Collaborative Peer Review and several exemplary collaboration scenarios that can take place during peer review. The next section introduces the ten courses in which Collaborative Peer Review was evaluated in and presents and discusses the results of the evaluations. Finally, the last section summarizes the chapter and discusses perspectives and implications for Collaborative Peer Review.

7.1 Communication during Peer Review

In his survey of peer review systems, Luxton-Reilly [LR09] states referring to dialogue in peer review systems that "[t]he systems considered here [in the article] vary substantially when it comes to supporting discussion within the peer assessment

framework" [LR09, p. 223] with the minority of examined systems providing communication means. The following section takes a closer look at the communication means of systems examined by Luxton-Reilly [LR09], as well as systems found in Søndergaard and Mulder's [SM12] survey of peer review systems, and other systems found in the scientific literature.

One of the earliest approaches at peer review using technology is described by Rada et al. [Rad+93] using the MUCH system. MUCH allows students to create documents as well as to access and assess all other students' documents. MUCH provides no means for communication but was used for peer assessment during a classroom session where students discussed documents and reviews after the review phase. Hence, dialogue was an element of the peer review process but not mediated by technology. Another early approach to peer review using technology is described by Downing and Brown [DB97] who provided students with a mailing list where they could send drafts to and receive feedback from their peers. As e-mail is an inherently bidirectional medium, Downing and Brown's approach enabled communication mediated by technology during the review phased.

Another, more recent approach which adopts face-to-face communication is described by Sitthiworachart and Joy [SJ03] where the review phase takes place during a lab session where students first review on their own and afterward discuss the reviewed works and their reviews in a group of three students. Afterward, outside the lab session, students are required to review the reviews they received for their work, that is, provide feedback to their reviewers. The approach of authors providing feedback to their reviewers (called *reverse review* by Wang et al. [Wan+16]) can be seen as a (simple) form of communication and can be found in various peer review systems, such as the system by Wang et al. [Wan+16], CrowdGrader [DAS14], PeerGrade [Gra17], Äropa [Ham+07], and SWoRD [CS07]. SWoRD, contrary to the other mentioned systems, includes two rounds of reverse review as students are required to turn in a revised version which is again subject to peer review, and consequently, the reviews to the revised version to reverse review [CS07]. Another form of reverse review is found in Expertiza where reviews are not reviewed by the author but by another participant completely unrelated to the reviewed work [Geh10].

The CAP system by Davies [Dav03] gives authors the possibility for anonymously contacting their reviewers for inquiries about the received reviews. The respective reviewer is then prompted to revise their review based on the request of the author. Such cycles of contacting reviewers and them revising their reviews can continue until the author is content with the received reviews. Similar feedback loops are supported by the OPAS system [Tra04].

Outright bidirectional communication between authors and each of their reviewers independently during the review phase is supported by Expertiza which allows addressing disagreements and inquiries as they arise [Geh10]. Similar functionality can be found in the system by Wang et al. [Wan+16] which provides a chat for the author and their reviewers. Äropa includes akin to the aforementioned systems the possibility for reverse review with the difference that this phase can optionally run concurrently to the review phase which allows reviewers to amend their reviews immediately in response to the author's feedback [Ham+07]. Maarek and McGregor [MM17] introduce a peer testing system where students provide feedback by writing unit tests for another students' source code. Their system allows the reviewer and author to chat after the tests have been run.

In Peer Grader, an earlier system by Gehringer [Geh01], authors and reviewers can communicate during the review phase, and reviewers of the same document can be given access to each other's reviews. Similar to the last aspect of Peer Grader, Hwang et al. [Hwa+08] used their collaborative annotation system VPen 2 for peer review where students had access to other students' annotations but no means for further communication.

In summary, existing peer review systems provide various means for communication: Starting with the possibility for authors to review their received reviews, sometimes combined with additional review cycles, to means where authors can contact a reviewer for resolving disagreements or addressing inquiries, to solutions that support bidirectional communication between author and reviewers. However, to the author's best knowledge, no system supports communication between all participants, that is, author, reviewers, as well as other participants unrelated to the review, at the same time. The next section introduces Collaborative Peer Review and its approach to bidirectional communication in more detail and provides exemplary scenarios for collaboration between the participants.

7.2 Collaborative Peer Review

Collaborative Peer Review enables novel collaboration and communication opportunities not possible in traditional peer review of writings. The following section outlines the characteristics of Collaborative Peer Review and explores possible collaboration scenarios that can take place in such an environment.

The cornerstone of Collaborative Peer Review is Backstage 2's collaborative annotation system: It allows participants to create their reviews using annotations placed at arbitrary locations of an essay, to comment on and up- and downvote annotations

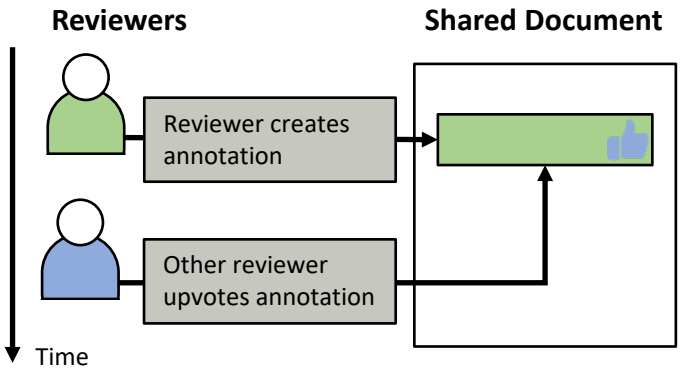


Fig. 7.1.: Collaboration between reviewers using the voting mechanism of the collaborative annotation system (Thumbs up icon made by Pixel perfect from `https://www.flaticon.com`).

and comments (see Chapter 3 for a detailed overview of Backstage 2's collaborative annotation system). Using annotations for reviews makes it easier to establish a context for a review, as reviewers no longer have to explain to which part of an essay a comment relates to [Bab+16] and allows "to-the-point comments" [Wan+16, p. 2016]. Note that the latter citation refers to peer review of program code which is similar to an essay as both are based on text.

Making essays and annotations immediately available to all participants of a course enables two things: First, communication between all participants, especially the stakeholders of the reviews, that is, authors and reviewers, and second, the opportunity for everyone to look at other participants' essays and reviews. As already discussed in the introduction, open access to all students' works is a rarely found feature in peer review systems.

The following paragraphs introduce possible communication and collaboration scenarios between authors and reviewers, between reviewers, and between reviewers and instructors.

Collaboration between Reviewers Among the possible outcomes when two or more reviewers create reviews for the same essay is that they create similar reviews for the same location, basically doing the same work more than once. Doing the same work more than once can be addressed by upvoting annotations – one reviewer can agree with another reviewer's annotation by just upvoting the respective annotation which is illustrated in Figure 7.1. Cho and Schunn [CS07] suggest that among the benefits of having more than one reviewer is that the same review given by more than one reviewer can be more convincing to the author. If an upvote is expressive enough to have a similar effect has to be shown.

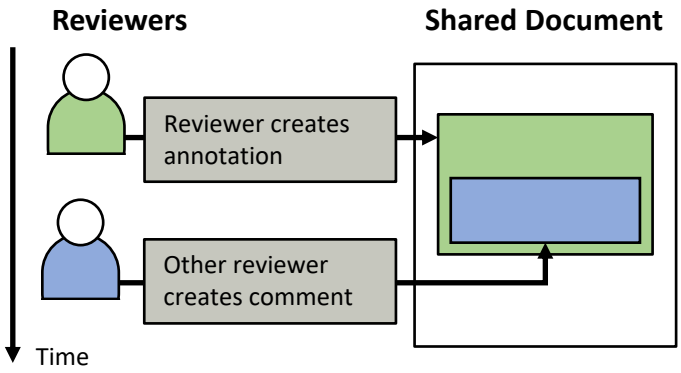


Fig. 7.2.: Collaboration between reviewers using comments.

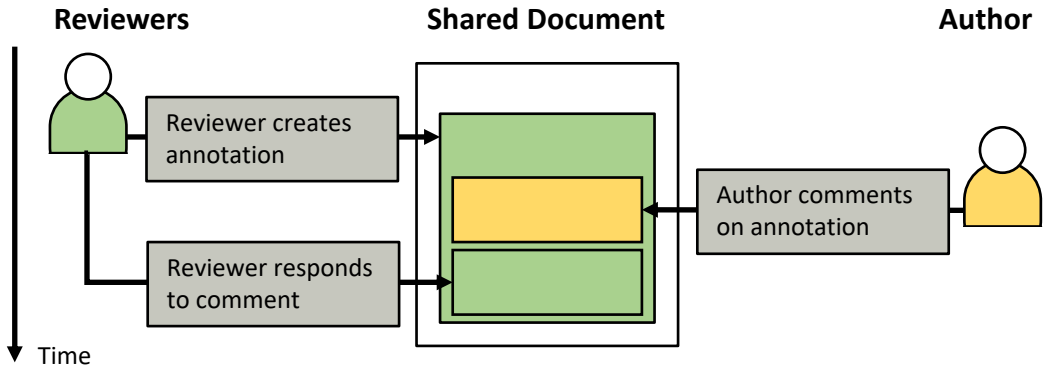


Fig. 7.3.: Collaboration between reviewers and authors using comments.

When voting is not enough, reviewers can use the comment functionality. Among the uses for the comment functionality are discussing disagreements between the reviewers that require resolution, agreeing with other reviewers but wanting to extend on the review, or just expressing a stronger agreement than possible through an upvote. Figure 7.2 illustrates this form of *active* collaboration between reviewers. While the figure shows only a single comment, it is obvious that conversations can span multiple comments as well.

Collaboration between Reviewers and Author In traditional peer review, authors have little to no opportunities to clarify aspects misunderstood by reviewers or inquire about received reviews after the review phase. With Collaborative Peer Review authors can comment on reviews while the review phase is still running, that is, while reviewers are still available, who, in turn, can revise their reviews or provide further explanations. Figure 7.3 illustrates a possible collaboration between reviewers and authors.

Collaboration between Reviewers and Instructor As discussed in Chapter 1, peer review can be utilized to lessen instructors' workloads and to provide a large number of students with formative feedback. In Collaborative Peer Review, instructors can still participate – with a reduced workload – in reviewing their students' essays, by

Tab. 7.1.: Overview of the course in which Collaborative Peer Review was used.

Course	Topic
Bachelor - high stakes	
CPR1	Web technologies
CPR2	Software design patterns
CPR4	Various computer science topics
CPR5	Various computer science topics
CPR9	Various computer science topics
CPR10	Web technologies
Bachelor - low stakes	
CPR3	Job applications
CPR6	Job applications
Master	
CPR7	Logic programming
CPR8	Computational ethics

acting as further reviewers who comment on and up- or downvote already existing reviews, and, if the need arises, create new reviews.

7.3 Study

Collaborative Peer Review was evaluated in ten courses, all of them being seminar-style courses with topics ranging from computer science topics to writing job applications. Table 7.1 shows an overview of the courses.

The courses were split into three groups: Courses targeting bachelor students and resulting in a numerical grade (*high stakes*), courses targeting bachelor students and not resulting in a numerical grade but either a pass or fail (*low stakes*), and courses targeting master students and resulting in a numerical grade. Distinguishing between those groups is important as they represent different levels of experience in writing essays (between master and bachelor students) as well as different levels of students' motivation (between high stakes and low stakes courses).

All courses followed a similar pattern: At the beginning of each course, participants were assigned a topic and tasked to research that topic on their own, write an essay, and create a presentation about the topic. At some point during the term, all participants were required to turn in preliminary versions of their essays which were then made available to all participants on Backstage 2. Except for **CPR8**, each essay was randomly assigned to two other participants for peer review. In **CPR8**, students worked in teams of two and every essay was assigned to two teams for peer

review, so every essay had two authors and four reviewers. As already discussed in Chapter 2, Backstage 2 has no formal assignment mechanism which means that the assignment of participants to essays for peer review was done manually outside of Backstage 2 by the instructors. The review phases ran for two to three weeks after which participants were given a few more weeks to revise their essays in response to the received reviews. In most of the courses, instructors provided their feedback using the collaborative annotation system as well. Furthermore, a few of the venues incorporated a final face-to-face session at the end of the term in which essays and reviews were discussed with all of the participants.

7.3.1 Methods

For evaluation, data was extracted directly from Backstage 2's database as well as collected through a survey conducted in six of the ten courses. The survey was identical in each course but was either conducted online after the course had concluded (in **CPR1-4**) or on paper during the last face-to-face session of the course (in **CPR5** and **CPR8**). The following section includes verbatim reproductions from [MB19d, p. 5f.] complemented with revisions and additions that reflect changes and additions to the evaluation.

Data Extraction from Database All annotations created by students for all essays were retrieved from the database and categorized as either *conversation annotation* or *review annotation*. *Conversation annotations* are annotations with at least one comment; conversely, *review annotations* are annotations without any comments.

Conversation annotations and their comments were further divided into *collaboration patterns* using the roles of their creators. To do that, each annotation and its comments were mapped to the role of the user who created that annotation or comment. For example, if an author commented on an annotation created by a reviewer, the resulting collaboration pattern is *reviewer-author*. Directly following comments of the same creator were conflated: For example, a reviewer commenting on their own annotation without anyone commenting in between has been assigned the collaboration pattern *reviewer*.

A note on determining the roles: As Backstage 2 has no formal assignment mechanism, the roles of participants in the context of an essay could not be retrieved from the database. One possibility would have been to manually map the authors and reviewers to essays, but that possibility was deemed unrealistic due to the associated effort. As a compromise solution, only authors were mapped manually to essays, and reviewers were determined using a heuristics: Those two (or four, for **CPR8**) participants who created most of the annotations for an essay and were not authors

of the essay were seen as reviewers. The heuristics was validated using a manually created mapping of essays to reviewers for the first three courses: Of the 38 essay-reviewers-pairs, 36 were identified correctly which is an acceptable correctness, and hence, the heuristics was deemed appropriate for identifying reviewers.

There was no small number of occurrences of the collaboration pattern *reviewer*, that is, the same reviewer commenting again on their own annotation without anyone commenting in between or at all on that annotation. In a previous evaluation (see [MB19d]), it was found that those comments were mostly used to edit annotations as the collaborative annotation system does not offer this functionality. Hence, these comments are omitted from the following evaluation, as they obviously do not represent collaboration.

Further classification of conversation annotations regarding their content was done for conversation annotations in **CPR1-3** as part of a previous study (see [MB19d]). Hence, the following description of the classification scheme is a reproduction of the respective part of [MB19d, p. 6] with the classification scheme for the pattern *reviewer* omitted as that pattern is not discussed in this chapter. Note that to avoid any possibility of confusion, the term *reviewee* (which was used in [MB19d] for the creator of an essay) was replaced in the following description by *author*.

To further examine what kind of communication took part in conversation annotations, those were classified after an original classification scheme by 3 judges ($\kappa = 0.59$, moderate agreement according to Landis and Koch [LK77]). The communication patterns *reviewee-reviewer*, *reviewer-reviewer-author*, and *reviewer-reviewer-reviewer* were omitted from the classification due to them appearing very rarely.

The communication pattern *reviewer-reviewer* was classified as follows:

- *agree*: The comment agrees with the review but does not extend upon it.
- *agree-extend*: The comment agrees with the review and extends upon it.
- *disagree*: The comment disagrees with the review but does not provide any justification for the disagreement.
- *disagree-extend*: The comment disagrees with the review and justifies the disagreement.

The communication pattern *reviewer-author* used the same classes as *reviewer-reviewer* extended with the following classes:

- *explanation*: The comment addresses misconceptions or answers a question.

- *inquiry*: The comment inquires about the review.

Furthermore, for each user-essay-pair, an *engagement duration*, that is, the time that the user spent on the respective essay, was calculated. The engagement duration was calculated using the following activity events:

- The event that was recorded to the database each time a participant accessed an essay.
- The event that was recorded each time a participant changed to another page of an essay.

For each user, all occurrences of those activity events were retrieved from the database, grouped by essay, and ordered chronologically. Subsequently, for the activity events that happened in the context of a single essay, time differences between directly following events were determined and differences larger than ten minutes removed. Summing up the remaining differences yielded the engagement duration for a participant for a respective essay.

Finally, all votes done on annotations and comments were retrieved from the database. Similar to the previously introduced collaboration patterns of annotations and their comments, collaboration patterns for votes were determined. In this case, these always included exactly two roles and represent the role of the voter and the role of the creator of the voted-on annotation or comment. For example, the pattern *reviewer-author* conforms to a vote done by an author on one of their reviewers' annotations or comments.

Survey The survey consisted, among others, of the following parts:

- A block of five items measuring the attitude towards giving peer review
- A block of five items measuring the attitude towards the received peer reviews
- A block of four items measuring the attitude towards the open access to all essays
- A block of four items measuring the attitude towards the open access to all peer reviews
- A block of four items measuring the attitude towards the course design

Answers to all those parts were given on a four-point Likert scale with no neutral choice in **CPR1-4** and on a six-point Likert scale with no neutral choice in **CPR5** and

Tab. 7.2.: Overview of the participants and average essay lengths in the examined courses.

Course	# of participants	# of pages (Mdn.)
CPR1	11	10.0
CPR2	13	11.0
CPR4	21	10.0
CPR5	18	11.5
CPR9	20	9.5
CPR10	11	12.0
CPR3	14	3.0
CPR6	14	2.5
CPR7	5	10.0
CPR8	10	11.0

CPR8. To make results comparable, results from **CPR1-4** were transformed linearly to the scale used in **CPR5** and **CPR8**, which ranged from *strongly agree* (assigned the value 5) to *strongly disagree* (assigned the value 0). That transformation led to identically named scale points not being assigned the same numerical value: Through the transformation, low values are generally overrated (*disagree* is transformed to a numerical value between *disagree* and *somewhat disagree*), and high values are generally underrated (*agree* is transformed to a numerical value between *somewhat agree* and *agree*).

Additionally, the survey contained items measuring the attitude towards Backstage 2, the System Usability Scale, three questions whether certain features were noticed by the participant, and three free text questions. As this chapter focusses on the collaborative aspects of peer review, evaluation of those questions was out of the scope of this chapter and therefore omitted. The entire survey can be found in Appendix A.3.

7.3.2 Results

Table 7.2 shows an overview of the number of participants and the median number of pages in each of the courses. Except for **CPR8**, the number of essays was equal to the number of participants; in **CPR8**, there were 5 essays, as participants worked in teams of two. Within the respective groups, essay lengths were comparable. Essays from the low stakes bachelor courses were, as expected, shorter, as job applications generally require less text than essays reporting on a scientific topic.

Fewer similarities are apparent in the annotation behavior of participants in each of the courses which can be seen in Table 7.3. Note that the absolute numbers of annotations are only reported for completeness' sake and are not suitable for

Tab. 7.3.: Overview of all annotations created during peer review.

Course	# of annotations	# of annotations per page	# of conversations	% of conversations
CPR1	664	6.0	21	3.2%
CPR2	451	3.1	41	9.1%
CPR4	787	3.4	34	4.3%
CPR5	468	2.4	12	2.6%
CPR9	602	2.9	14	2.3%
CPR10	412	3.2	26	6.3%
CPR3	219	4.2	6	2.7%
CPR6	243	6.4	12	4.9%
CPR7	78	1.5	12	15.4%
CPR8	424	6.3	40	9.4%

comparing courses that vary in the number of participants and the number of pages to be reviewed. Annotations done per page, used as a crude measure of review activity, varied greatly across the courses as well as within each of the three groups with no apparent trend. The relative number of conversation annotations exhibited no trend as well. For the high stakes bachelor courses, it can be argued the **CPR2** is an outlier and the percentages between 2.3% and 6.3% observed in the other courses are more of the norm for that group. The master courses showed compared to the other courses a higher percentage of conversation annotations. Note that this observation could be a coincidence due to the low number of examined courses in that group.

The following section first examines first conversation annotations in more detail before looking at the usage of voting for passive collaboration. The engagement duration and the engagement of participants in general are reported in the part after that. The final part of this section presents the students' attitudes towards peer review, open access to essays and reviews, and the course design in general.

Active Collaboration Patterns Looking in more detail at conversation annotations, annotations with up to three comments (conforming to a communication length of 4 when including the initial annotation) were observed. As already mentioned in Section 7.3.1, the only observed pattern with a communication length of 1, *reviewer*, is omitted, as the usages of that pattern did not represent collaboration.

Around 92% of all annotations had a communication length of 2 which conforms to annotations with a single comment from another user. Within those annotations with communication length 2, only the patterns *reviewer-author* and *reviewer-reviewer* were observed more than twice. An overview of the occurrence of those patterns

Tab. 7.4.: Overview of collaboration pattern with a communication length of 2 across all courses. The given percentage values are relative to all conversation annotations and not only those with a communication length of 2.

Course	*reviewer-reviewer*		*reviewer-author*	
	absolute	relative	absolute	relative
CPR1	8	38.0%	11	52.4%
CPR2	24	58.5%	16	39.0%
CPR4	8	23.5%	21	61.8%
CPR5	3	25.0%	7	58.3%
CPR9	9	64.3%	4	28.6%
CPR10	4	15.4%	18	69.2%
CPR3	2	33.3%	4	66.7%
CPR6	3	25.0%	5	41.7%
CPR7	3	25.0%	6	50.0%
CPR8	4	10.0%	35	87.5%
Percentage	31.2%		58.3%	

across all courses can be seen in Table 7.4. While there is no consistent picture, with two exceptions, the pattern *reviewer-author* was more prevalent. Indeed, when taking all courses together, *reviewer-author* was occurring twice as often as *reviewer-reviewer*.

Communication lengths greater than two were rare; conversation annotations with communication length 3 made up only around 7% of all conversation annotations. Within those conversation annotations, only the patterns *reviewer-reviewer-author* (2.8% of all conversation annotations) and *reviewer-author-reviewer* (2.3% of all conversation annotations) occurred more than twice. Longer communication lengths were nearly never observed with only *reviewer-reviewer-author-reviewer* being observed more than once across all courses.

Table 7.5 shows the classification of conversation annotations with communication length 2 by their content for **CPR1-3** done in [MB19d]. Classes in italics are those classes that introduce new aspects into conversations or would require further action by other conversation participants. Those kinds of annotations made up 70% of all conversation annotations with communication length 2, that is, have a clear majority in the examined sample.

Passive Collaboration Patterns Coming to passive collaboration, that is, collaboration by voting on others' annotations and comments, a total of 1408 votes were done across all courses. The number of votes and votes per annotation broken down by courses can be seen in Table 7.6. Note that the relative measure was obtained

Tab. 7.5.: Classification of conversation annotations with communication length 2 by their content (taken from [MB19d, p. 7], removed pattern *reviewer*, replaced *reviewee* with *author*).

Class	CPR1	CPR2	CPR3
reviewer-reviewer			
agree	0	5	0
disagree	2	1	1
agree-extend	4	10	0
disagree-extend	1	6	1
reviewer-author			
agree	4	0	1
disagree	0	1	0
agree-extends	2	2	0
disagree-extends	2	7	1
explanation	1	6	0
inquire	2	0	2
miscellaneous	1	4	0

by dividing through the number of annotations which disregards the number of comments. While including the comments would decrease the relative measures slightly, it would not have a significant effect on the magnitude and differences between the courses.

Similar to the measures reported until now, the numbers varied greatly across the courses with no apparent trend. Across all courses, roughly every third annotation was voted on (Mdn: 0.31, MAD: 0.14), but that number fluctuates greatly: There are courses where every second annotation received a vote (**CPR1** and **CPR10**), but also courses were only every thirtieth annotation received a vote (**CPR3**).

Looking at the patterns of passive collaboration, that is, the role of the voter and the role of the creator of the voted-on annotation or comment, ten different patterns emerged. Of those ten patterns, four were observed in at least half of the courses.

Most often occurred the pattern *reviewer-author*, that is, an author voting on one of their reviewers' annotations which made up 54.4% of all votes.

The second next pattern was *reviewer-reviewer* with 33.6% of all votes. Hence, what was already observed for active collaboration, that is, that collaboration between reviewers and authors happened more often than other forms of collaboration, can be extended to passive collaboration as well. For the other patterns, which occurred considerably less often, in the majority of courses, the patterns *reviewer-other* and *reviewer* (4.3% and 4.5% of all votes, respectively) were observed. The former

Tab. 7.6.: Overview of the votes done and the average votes per annotation across all courses.

Course	# of votes	Votes per annotation
CPR1	317	0.48
CPR2	123	0.27
CPR4	233	0.30
CPR5	197	0.42
CPR9	132	0.22
CPR10	202	0.49
CPR3	7	0.03
CPR6	79	0.33
CPR7	27	0.35
CPR8	91	0.21

is someone who has no role in the context of an essay voting on an annotation or comment referring to that essay; the latter is a reviewer voting on their own annotation or comment.

Students' Engagement The engagement was determined in two ways: First, as the total time participants spent in their various roles, that is, the time reviewers spent on essays assigned to them for review, the time authors spent on their own essays, and finally, the time participants spent on essays they were neither assigned for review nor the author of. The latter number was used to determine the number of essays a user engaged with besides those they had a stake in (i.e., either as reviewer or author): Both the number of such essays a user engaged with regardless of the time spent as well as the number of such essays a user viewed *meaningfully* were determined from that number. Meaningful was interpreted very liberal in this evaluation: An essay was seen as having been viewed meaningfully by a user if it was viewed for more than one minute by that user.

The durations participants spent in their respective roles per essay can be seen in Table 7.7. Not taking into account **CPR3** and **CPR6** (as essays in those courses were much shorter compared to the essays in the rest of the courses), reviewers spent in median around an hour for reviewing an essay (Mdn: 57.8), and authors around three-quarters of an hour on their essays (Mdn: 45.4).

The number of essays participants viewed *besides* those they had a stake in (i.e., either as reviewer or author) can be seen in Table 7.8 both in absolute numbers as well as relative to the total number of essays (minus those they had a role in). Across all courses, participants viewed other participants' essays, but the percentage of viewed essays varied greatly across the venues. As a general trend, participants

Tab. 7.7.: Time spent by participants for the respective task per essay across all courses.

Course	Median time spent by reviewers in minutes	Median time spent by authors in minutes
CPR1	67.7	47.1
CPR2	64.1	43.7
CPR4	68.2	42.2
CPR5	40.1	43.5
CPR9	51.3	48.9
CPR10	79.7	63.8
CPR3	27.6	15.9
CPR6	42.5	21.6
CPR7	40.6	23.6
CPR8	51.5	59.4

Tab. 7.8.: Number of essays students spent viewing regardless of the time spent and number of essays viewed meaningfully (i.e., longer than one minute) by students.

Course	Median # of viewed essays		Median # of meaningfully viewed essays	
	abs.	rel.	abs.	rel.
CPR1	4	50.0%	2	25.0%
CPR2	7	70.0%	3	30.0%
CPR4	7	38.9%	3	16.7%
CPR5	6.5	43.3%	2	13.3%
CPR9	9.5	55.9%	2.5	14.7%
CPR10	1.0	12.5%	1	12.5%
CPR3	10.5	95.5%	7.5	68.1%
CPR6	11	100.0%	11	100.0%
CPR7	2	100.0%	2	100.0%
CPR8	0.5	25.0%	0	0.0%

did not engage with other essays for a longer time, with participants generally viewing two other essays meaningfully (Mdn: 2.25). Outliers are **CPR3** and **CPR6** in which the majority of essays were viewed meaningfully by at least half of the students. **CPR3** and **CPR6** had in common that essays were shorter and pertained with writing job applications a soft skills topic as opposed to scientific writing in the other courses. Complete opposites were displayed by the master courses: While in **CPR7** the majority of participants spent more than one minute with all of the essays, in **CPR8** participants exhibited nearly no engagement outside of the essays they had a role in.

Students' Attitudes In **CPR1-5** and **CPR8**, students' attitudes towards peer review and the peer review environment were measured using a survey. As already discussed,

the survey used a four-point Likert scale in **CPR1-4**, and a six-point Likert scale in **CPR5** and **CPR8**. The results from **CPR1-4** were linearly transformed to the scale used in the other courses which led to low values being overrated and high values being underrated.

The results on giving and receiving peer review in general (i.e., without any mention of the open access to essays and reviews) can be seen in Table 7.9. Across all courses, students valued the various aspects of peer review. Students agreed that both giving peer review and the received peer reviews helped them to improve their essay and their writing in general. Giving peer review enabled students to get an understanding of the standard of work in the course and helped them to assess their own performance. Only in **CPR2**, students were undecided on the positive effects of their received reviews on their work and writing. In any case, across all courses, students disagreed with the negatively phrased items asking whether giving peer review had few to none positive aspects and whether the received reviews had little to no use to them, which suggests that students across all courses benefited somehow from peer review.

On the question, whether the received peer reviews were more valuable than lecturers' feedback, students were divided: There are three courses where students tended to agree (**CPR1**, **CPR4**, and **CPR8**), one course where students were undecided (**CPR3**), and two courses where students tended to disagree (**CPR2** and **CPR5**).

Regarding the attitudes towards open access to essays and reviews, the aggregated students' responses can be seen in Table 7.10. Note that those items were only answered by those students who had previously confirmed that they viewed other participants' essays (besides those assigned to them) or reviews. In **CPR1-4** that was done through a conditional question in the online survey; in **CPR5** and **CPR8** through an introduction statement above the respective blocks in the survey.

Students' attitudes towards open access to essays and reviews were generally positive and students strongly disagreed with the statements that access to either had no positive effects on their own essay. For open access to reviews, students found that aspect to have positive effects on the quality of their own essay. Furthermore, in most courses, students agreed that access to other participants' reviews gave them ideas for their own reviews and that they used suggestions made for other essays to improve their own essays. Results referring to the open access to essays indicate that this aspect can promote similar effects than doing peer review as students thought that this aspect helped them to better assess their performance and to get a feeling for the standard of work in the course. Furthermore, students mostly agreed that they found aspects in other participants' essays that helped them improve their own essays.

Tab. 7.9.: Aggregated students' responses to the items measuring the attitude towards giving peer review and the received peer reviews. Items marked with (*) were phrased negatively in the survey (shortened items adapted from [MB19d, p. 8]).

Statement	CPR1 n = 8 Mdn	CPR2 n = 6 Mdn	CPR3 n = 4 Mdn	CPR4 n = 5 Mdn	CPR5 n = 15 Mdn	CPR 8 n = 8 Mdn
Giving peer review						
New ideas to improve essay	3.3	3.3	3.3	3.3	4.0	4.0
Better understanding of standard of work in course (*)	5.0	3.3	3.3	3.3	4.0	4.0
Beneficial to learning of writing	3.3	3.3	4.2	3.3	4.0	3.0
Compared to other courses a better assessment performance	3.3	3.3	3.3	3.3	4.0	4.0
Few to none positive aspects	0.0	1.7	0.0	1.7	1.0	1.0
Received peer reviews						
Helped to greatly improve essay	4.2	2.5	3.3	3.3	4.0	4.0
Beneficial to learning of writing (*)	5.0	3.3	4.2	3.3	4.0	4.0
Opened up new perspectives on writing essays	3.3	2.5	4.2	3.3	3.0	3.0
Little to no use	0.0	1.7	0.8	1.7	1.0	1.0
More valuable than lecturers' feedback	3.3	1.7	2.5	3.3	2.0	2.8

Tab. 7.10.: Aggregated students' responses to the items measuring the attitude towards the open access to essays and reviews.

Statement	CPR1 n = 8 Mdn	CPR2 n = 6 Mdn	CPR 3 n = 4 Mdn	CPR4 n = 5 Mdn	CPR5 n = 11-13 Mdn	CPR 8 n = 5 Mdn
Access to all essays						
Helped assess own performance	3.3	3.3	3.3	3.3	4.0	4.0
Used aspects found other essays to improve own essay	3.3	2.5	3.3	3.3	3.0	2.0
Feeling for the standard of work in the course	3.3	3.3	3.3	3.3	5.0	3.0
Little to no positive effects on my essay	0.0	1.7	0.0	1.7	1.0	1.0
Access to all reviews						
Ideas for my own peer review	3.3	3.3	1.7	3.3	4.0	4.0
Used suggestions made for other essays to improve own essay	2.5	3.3	3.3	3.3	3.0	2.0
Positive effects on the quality of own essay	3.3	3.3	3.3	3.3	4.0	4.0
No positive effects on my essay	1.7	0.0	1.7	1.7	1.0	1.0

Tab. 7.11.: Aggregated students' responses to the items measuring the attitude towards the course design.

Statement	CPR1 n = 8 Mdn	CPR2 n = 6 Mdn	CPR 3 n = 4 Mdn	CPR4 n = 5 Mdn	CPR5 n = 14 Mdn	CPR 8 n = 8 Mdn
I think that peer review was a good fit for the course.	5.0	3.3	5.0	3.3	4.0	4.0
Giving peer review was too time-consuming.	1.7	1.7	0.8	1.7	2.0	1.0
I would have preferred a more traditional course design.	0.8	1.7	0.0	1.7	1.0	1.0
Peer reviews from a single reviewer would have been sufficient.	1.7	0.8	0.0	1.7	1.0	1.0

The final block of survey measured the students' attitudes towards the course design. The aggregated students' responses to those items can be seen in Table 7.11. Across all courses, students agreed with the sentiment that peer review was a good fit for the course. Additionally, students strongly preferred the course design to a traditional course design, that is, a course without peer review. Furthermore, students found giving peer review not to be too time-consuming and disagreed that a single reviewer would have been sufficient, that is, preferred to have more than one reviewer for their essays.

7.3.3 Discussion

Collaborative Peer Review is a learning format that enables participants to collaborate during the review phase by giving every participant access to all essays and reviews as they are created. Collaboration can take place in two ways: Actively, by creating and commenting on annotations, and passively, through voting on annotations and comments.

Active collaboration most often ended after a single comment, that is, most of the time a comment remained without answer. While there are cases in which a single comment concludes a conversation, such as a comment that agrees or extends upon a review, there are cases that warrant further reaction by other participants, such as a comment disagreeing with or inquiring about the review. If the results regarding the content of conversation annotations done for **CPR1-3** can be generalized (which is likely due to the similarities of the courses), that would mean that conversations

often end too early. Hence, means for nudging participants to resolve conversations have to be devised.

On a positive note, the analysis of the contents of conversation annotations also revealed that in the majority of cases, a comment added something to a review, that is, brought additional value to the review. Furthermore, there are several (unfortunately unresolved) disagreements where one reviewer disagrees with another reviewer's annotation and provides reasons for the disagreement which is impossible in a traditional peer review environment. Even unresolved, those disagreements present a valuable resource for the author, because they are presented another perspective on the review.

Throughout the majority of courses, participants used the voting functionality quite extensively. The most commonly observed pattern of an author voting on a reviewers' annotation was not foreseen as a possible pattern and does not really represent a form of collaboration. It is imaginable that authors used those votes to either show their agreement or disagreement with a review or as a kind of bookmark to mark those reviews which were already taken notice of. Otherwise, a large number of votes where one reviewer votes on another reviewers' annotation or comment shows that reviewers used that feature most likely as suggested at the beginning of this chapter. If a user participated in an essay there were neither assigned for review nor the author of, that participation took the form of voting on annotations and comments; users creating annotations for or commenting on annotations referring to essays they had no role in was nearly never observed. A likely explanation for that behavior is the difference in the effort between creating an annotation or comment and a vote which only requires a single click on a button. Nonetheless, the mere existence of votes from participants unrelated to the essay shows that the open environment opens up new ways of collaboration during peer review.

The majority of collaboration, regardless of active or passive, took place between reviewers and authors which can be taken as an indication that students still saw the peer review process as something organized into phases and no as a continuous process spanning a longer time period. As authors are generally expected to work with the reviews after the review phase has concluded, they view essays after all reviews have been created, that is, have the opportunity to interact with all created reviews. With reviewers, on the other hand, it is possible that they finish their reviews and never return, that is, such reviewers have only the opportunity to interact with those reviews that were created by the time they created their reviews.

Depending on the course, students engaged differently and different amounts of time with the essays. Looking only at courses with essays of similar length (i.e., all courses except **CPR3** and **CPR6**), authors spent in median around three-quarters

hours on their essays and reviewers around an hour for reviewing each of their assigned essays. Using time spent as a rough estimation of work done indicates that reviewers spent an appropriate amount of time for reviewing and that peer review increased the time on the task (which is suggested by Topping [Top98] as one of the benefits of peer review). Authors working around 45 minutes with the received reviews suggests that they received a non-negligible amount of useful feedback. Looking at the number of essays that participants engaged longer than a minute with, participants engaged in median with two essays besides those they had a role in. While that number might seem low at first, those are two more essays than the same participant would have engaged with in a traditional peer review environment.

Outliers from that rule are courses **CPR3** and **CPR6** where the majority of students engaged more than one minute with the majority of essays. In contrast to the other courses, those courses covered the same soft skills topic, namely writing job applications. Hence, the essays in those courses were job applications written by the participants to fake job adverts as opposed to scientific essays in the other courses. That difference suggests possible explanations for the participants' differing behavior: Students are more inclined to browse an essay of around 3 pages compared to an essay of around 10 pages, less buy-in is required for reading a non-scientific essay compared to a scientific essay, or students find more value in reading other participants' applications. Indeed, it might be easier to learn from another person's (well-written) application than from another person's scientific essay. Other outliers are the two master courses in regard to the relative number of conversation annotations: In both courses, considerably more annotations than in the other courses were commented on. For Collaborative Peer Review these observations suggest that in different contexts and for different topics, different levels of student engagement and collaboration are to be expected.

Considering the results of the survey, Collaborative Peer Review can be considered a success: Students benefitted from giving peer review and the received peer reviews. Note that those results are not exclusive to Collaborative Peer Review, and similar results are to be expected in a traditional peer review environment. What can be attributed to Collaborative Peer Review are the positive results referring to the open access to essays and reviews where results suggest (with few outliers) that the open environment supports students in assessing their performance and getting an overview of the standard of work in the course. Furthermore, the open access might provide them cues on how to create their own reviews and improve their own essay.

This section closes the evaluation of Collaborative Peer Review. The evaluation found evidence that adding collaboration to peer review can enhance reviews and

promote communication and collaboration between the different stakeholders of a review. Nonetheless, to transition Collaborative Peer Review from an observation to a veritable learning format, further improvements have to be made which are outlined in the following section.

7.4 Wrapping up Collaborative Peer Review

Collaborative Peer Review was created from observations: In the first venues (**CPR1-3** and to some extent the course discussed in [MB18a]), students were found to use the comment and vote functionalities of the collaborative annotation system during peer review. Providing means for communication and collaboration during the review phase addresses various issues of traditional peer review: Authors are provided opportunities to inquire about reviews before the end of the review phase, and reviewers can collaborate to resolve disagreements and are, through having access to the other reviewers' reviews, prevented from doing the same work twice.

First things first, even in an environment that was not specifically built for peer review, students found the reviewing and the received reviews to help them improve their essays and spent an appropriate amount of time creating and working with the reviews. Furthermore, students found open access to essays and reviews to have positive effects on their essays and self-assessment.

One of the shortcomings identified in the evaluation was that conversations most often remained unresolved, such as a reviewer disagreeing with another reviewer's review which is a situation where one would expect the initial reviewer to either argue for or adapt their review. Unresolved conversations might be an effect of insufficient communication awareness: If participants are not aware that something happened, they cannot react to that. As already discussed, Backstage 2 has no explicit means for communication awareness on that level which means that participants had to completely browse an essay to get an overview of new annotations and comments.

Among the possible approaches for promoting communication awareness are interface cues and an overview of activity not already seen by a user. Nikolai Gruschke developed two approaches for communication awareness as part of his unpublished bachelor thesis which can be seen in Figures 7.4 and 7.5.

A widget for Backstage 2's dashboard (see Section 2.2) was developed which can be seen in Figure 7.4. That widget informs a user about new comments on that user's essay or reactions on annotations or comments created by that user.

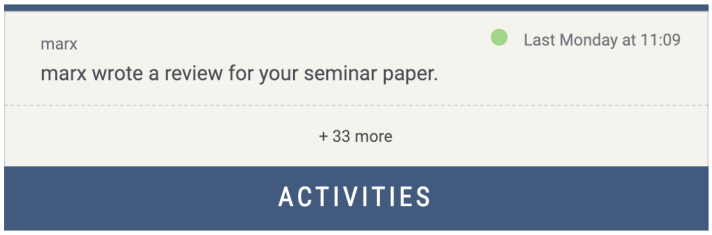


Fig. 7.4.: Dashboard notifying a user about a new review for the user's essay.

Fig. 7.5.: Interface element allowing a user to browse an essay. A green dot about a number indicates a page with unseen activity (taken from [MB19d, p. 8]).

Figure 7.5 shows another mechanism for communication awareness in form of an updated pagination, which is shown in the detail view of a unit (see Section 2.3), where a green dot about a page number indicates unseen activity on that page. Evaluation of the effects of these approaches on Collaborative Peer Review was planned starting from **CPR4** and indeed available during the peer review phase in **CPR4**, but, due to bugs, correct working during **CPR4** could not be guaranteed. A lack of time prevented those bugs from being fixed, and hence, the approaches to communication awareness remain unevaluated.

While not as important as communication awareness, another kind of awareness could be important for Collaborative Peer Review as well: Task awareness. As already discussed, Backstage 2 has no formal assignment mechanism, and hence, the system is not aware of who reviews what and, therefore, cannot estimate participants' progress in their assigned reviews. Having such an estimation would allow Backstage 2 to give participants feedback on the completion of their reviews as well as instructors an overview of all participants' progress. As previously discussed, estimating the progress is not an easy task but a few of the measures and their observed characteristics introduced in the evaluation, such as the engagement duration or the number of annotations, could be used. However, showing whether the system considers reviews as finished could have detrimental effects as well: Participants stopping to review as soon as the system reports their reviews as done which could intensify the effect of reviewers collaborating more rarely than reviewers and authors.

There are avenues that should be considered for future research: First, having all reviewers work on the same copy of an essay might introduce bias with the reviewer who created their reviews first implicitly guiding the reviews of the following reviewers. Similarly, reviewers could be inclined to not disagree with the other reviewers, which might lead to them silently agreeing through inaction even though

they actually disagree with a review. One assumption of Collaborative Peer Review stated at the beginning of this chapter was that voting has a similar effect on the perceived importance of a review than a further review stating the same. That assumption was not evaluated as part of this study but should be considered for future research. Finally, one aspect completely omitted from the evaluation is the quality of the reviews which is an important aspect of peer review. Quality was not evaluated as assessing a large number of formative reviews is a major effort that could not be done as part of this thesis. Anecdotally, in the first three venues (which were supervised by the author), the quality of reviews varied but was mostly acceptable and made providing formative feedback less time-consuming.

Limitations to the evaluations described in this chapter pertain to the comparability of the courses: The courses were held by different lecturers, and hence, different instructions on how the peer review should be approached have been given which might have influenced the way participants did their peer review and engaged in collaboration. Furthermore, the courses covered different topics and it might be possible that some topics lend themselves more to a collaborative approach to peer review than others.

This closes the chapter on Collaborative Peer Review, a format where instructors take a further step back, transitioning from a facilitator of learning to an orchestrator of learning. In the next format, Bite-sized Learning, the instructor nearly completely vanishes: Instructors only create quizzes and learning material, and students learn self-paced on their own supported by technology.

8 Bite-sized Learning

In the learning formats introduced in the previous chapters, there was always accompanying face-to-face teaching and the use of Backstage 2 was mostly orchestrated by lecturers. This final chapter of Part II introduces a format where Backstage 2 is not used in conjunction with face-to-face teaching, but to provide students additional learning opportunities to work on at their own pace. These opportunities are divided into *bite-sized* units to make it easier to work through them, hence the name of the format.

In Bite-sized Learning, a *bite* is a collection of learning resources learners can work through in a short amount of time. Learning resources are generally quizzes but can be other material, such as videos or short texts, as well. Hence, the audience response system is the only component required for implementing Bite-sized Learning in Backstage 2.

Bite-sized Learning was evaluated in two courses in two different fields: A course on medicine, and a course on Ancient Egypt. The course on medicine was available the week before the examination as an additional opportunity for examination preparation; the course on Ancient Egypt was available throughout the term as an opportunity for students to catch-up on knowledge.

Note that Colin Gray [Gra15] defines as part of his doctoral thesis a bite-sized learning format as well, which, while sharing similarities with the format introduced in this chapter, also includes restrictions and conditions, such as requiring social interaction between the participants, that do not conform with the understanding of Bite-sized Learning represented in this chapter.

This chapter first discusses Microlearning, which is the foundation of Bite-sized Learning, its characteristics, effects on students' learning, and exemplary applications. Then, first, the course on medicine and the results of its evaluation are presented and discussed. After that, the same is done for the evaluation in the course on Ancient Egypt. Finally, the results of both courses are compared, conclusions for Bite-sized Learning are drawn, and future research directions are discussed.

8.1 Microlearning

Microlearning is "a learner's short interaction with a learning matter broken down to very small bits" [Lin06, p. 46], "learning in tiny chunks and short bursts of time" [Jom+16, p. 103], or "fine-grained, interconnected but loosely coupled learning opportunities" [Sch07, p. 99]. While all those definitions are phrased differently, they share the same core – learning takes place in "small bits", "tiny chunks", or "fine-grained", that is, in forms that can be consumed by a learner in a relatively short amount of time. Another term for that concept often used in the context of microlearning is "bite-size" (see, e.g., [ZW19; Giu17; Pou13]).

For the duration of a *bite*, various figures can be found in the literature: Lindner [Lin06] mentions durations from a few seconds to 15 minutes, Zhang and West [ZW19] state that a microlearning session should take no longer than 20 minutes to complete. Similarly, Alqurashi [Alq17] takes the view that a microlearning session should be able to be completed in 15 to 20 minutes. Without providing references, Torgerson [Tor16] states that opinions for the duration of a microlearning session range from 90 seconds to 10 minutes, but personally believes that a session should last no more than 5 minutes. In summary, *bites* that can be completed in under 20 minutes can be considered microlearning sessions.

Note that microlearning is not limited to formal educational settings, but covers every situation in which engagement with a learning resource takes place for a short duration [Sch07]. An example of microlearning outside educational settings is Cai et al.'s [Cai+15] *Wait-Learning* in which users answer multiple choice questions while they are waiting for an answer in a conversation using a text messenger. Similarly, Kovacs [Kov15] shows multiple choice questions in a user's Facebook newsfeed. Furthermore, platforms such as Duolingo[1] or Khan Academy[2] can be considered microlearning platforms. Even video platforms such as TED[3] or YouTube[4] can be seen as microlearning platforms, as their users create among others content that can be used for short sessions of learning.

As the examples above make evident, microlearning is often occurring in the context of e-learning platforms or at least online, with Lindner [Lin06] even restricting his definition to e-learning. Based on that, everything available to e-learning platforms can be used to create microlearning sessions. Alqurashi [Alq17] mentions static learning resources, such as videos or podcasts followed by interactive learning resources in the form of quizzes. Similarly, Nikou [Nik19] lists static learning

[1] https://www.duolingo.com/
[2] https://www.khanacademy.org/
[3] https://www.ted.com/
[4] https://www.youtube.com/

resources and interactive learning resources, such as formative assessment with feedback, but mentions collaborative activities, such as peer review, as well.

Interactive learning resources, such as quizzes, provide opportunities to make microlearning sessions more active and are associated with higher learning achievement compared to simply viewing static contents [Koe+15]. Furthermore, quizzes are a form of tests, and that is where the testing effect comes into play: The testing effect describes the phenomenon that "[t]aking a test on material can have a greater positive effect on future retention of that material than spending an equivalent amount of time restudying the material" [RIK06, p. 181]. In their meta-survey, Roedinger and Karpicke [RIK06] provide an extensive overview of the literature on the testing effect which mostly agrees that being tested results in better learning achievement compared to not being tested at all or simply studying.

Kibble [Kib07] provided sessions of about 20 to 30 multiple choice quizzes to students before examinations where the quizzes were of similar difficulty to those in the examination. They provided the same quizzes in five consecutive terms and added each term more reward for doing the quizzes in the form of course credit. Their results show that doing those quizzes had a positive effect on examination performance and that the rewards had a positive effect on the number of students who participated in the online quizzes.

Johnson and Kiviniemi [JK09] tasked students to answer multiple choice quizzes referring to a weekly reading assignment. Their system presented the quizzes in sessions of 10 quizzes, which were randomly selected from a pool of 25 quizzes, and only if all of the quizzes in a session were answered correctly, the quizzes for the respective reading assignment were counted as completed. Their results show that the more quiz sessions were completed by a user, the higher their score in the examinations was. Furthermore, the authors explicitly eliminate the possibility of the observation just being an effect of only well-performing students doing the quizzes by showing that no correlation between quiz session completion and performance on parts on the examination not tested by the quizzes existed.

Angus and Watson [AW09] employed four online quiz sessions throughout the term where successful completion of each rewarded course credit. Their results show that students who attempted all quiz sessions (compared to students who did attempt three or less) had significantly better examination results.

The positive effect of quizzes makes them an attractive element for designing Bite-sized Learning. Therefore, in the following, two courses utilizing Backstage 2 for Bite-sized Learning are introduced. Both courses consist exclusively of quizzes which

provide feedback on an answer's correctness, on the performance of one's peers, and in the form of texts that explain the correct answer.

8.2 Examination Preparation Course for Medicine

Bite-sized Learning was evaluated first in a course on medicine, specifically neurology. The course was provided to students as an additional opportunity for examination preparation the week before the examination. Due to a high number of students, the examination is not written by all students on the same day; instead, students are split into two groups, each group with their own examination date. The first group wrote their examination in December 2017, the second group in February 2017. Both groups were provided the same course on Backstage 2 the week before their examination.

The course consisted of 90 quizzes which were handcrafted by Franz Pfister, Konstantin Dimitriadis, and Boj Hoppe. These quizzes were split into 6 sessions containing 15 quizzes each. To ensure variety, each session consisted of a mix of multiple choice, open answer, scale, and mark the region quizzes. The majority of quizzes consisted of a description of a medical case which was sometimes accompanied by an image or video, followed by a quiz referring to that case. After giving an answer, students were shown an explanation of the correct answer as well as an overview of their peers' performance on that quiz. As an incentive for students to do the quizzes, five images used in the quizzes were also used in examination questions which was disclosed beforehand.

Examples for multiple choice quizzes include deciding on which medication to prescribe or on the most likely diagnosis. Open answer quizzes asked similar questions but users had to write their answers instead of choosing from provided answer options. Scale quizzes were used when the answer was a number, such as determining the score on the Glasgow Coma Scale (see [Ste16]) for a medical case.

For mark the region quizzes, radiological images were used with the task in the majority of cases being to mark the pathological region as a polygonal selection. A screenshot for a mark the region quiz can be seen in Figure 8.1. The screenshot shows the quiz after an answer has been given: The red polygon represents the given answer, the blue polygon the correct answer. Due to the answer being incorrect (as the red polygon is not at the same or a similar position as the blue polygon), the polygon is colored red; a correct answer would lead to a green polygon. On the left, there is further feedback on the answers' correctness, as well as the aforementioned

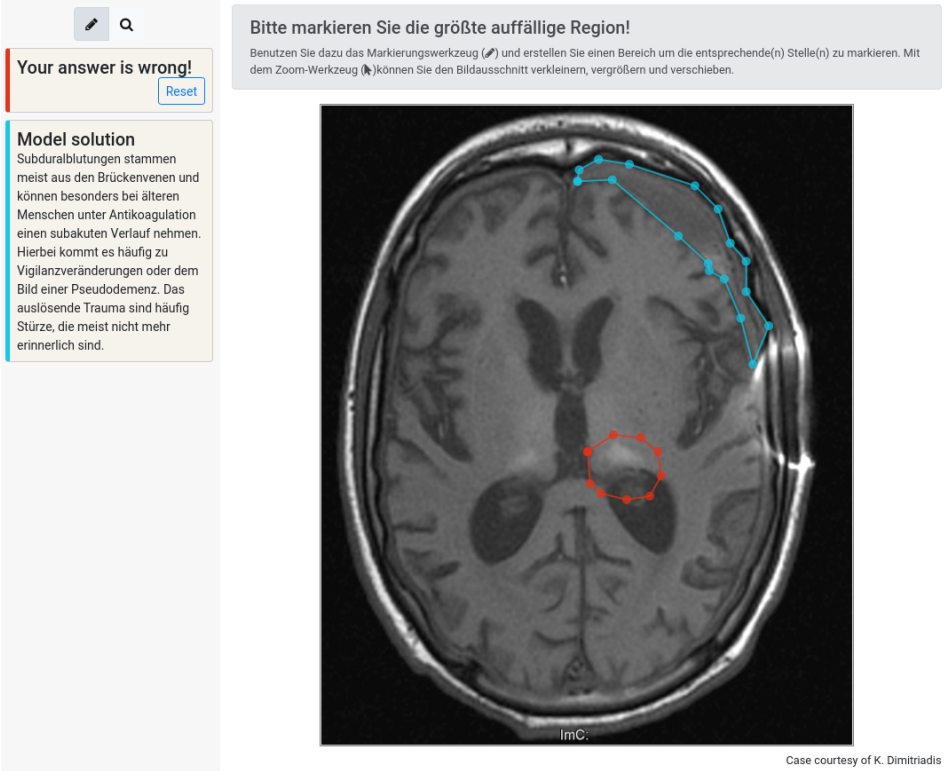


Fig. 8.1.: Screenshot of a mark the region quiz: The red polygon shows the user's answer; the blue polygon the correct answer. On the left, there is further correctness feedback as well as an explanation of the image (annotated image copyright of Konstantinos Dimitriadis).

explanation text which describes what is wrong in the image. As already mentioned, due to time constraints, mark the region was not implemented as a question type in the audience response system, but rather using the collaborative annotation system. Therefore, mark the region quizzes did not give an overview of the other participants' correctness.

For the second group of students, no major changes were made: Errors in quizzes were fixed, and the answers to open answer quizzes were evaluated more leniently so that answers with one or two scrambled or missing letters were still accepted as the correct answer.

The remainder of this section describes the evaluation of the course on medicine in two venues, **M1** and **M2**, and discusses the results.

8.2.1 Methods

For evaluation, data from two sources was used: Data collected directly from Backstage 2, and data from surveys conducted online after the respective examination of the group.

Not all parts of the survey are evaluated in the evaluation described below. The following list only mentions those parts which were used in the evaluation. The entire survey can be found in Appendix A.4.

1. Four questions asking for each question type how helpful the respective type was for examination preparation.
2. Four questions asking for each question type how clear the process of submitting an answer was.
3. Four questions asking for each question type how understandable the correctness feedback was.
4. One question asking if the comparison with their peer's supported them in assessing their current learning progress, followed by a question if they would have liked more comparison with their peers.
5. A block of questions measuring the construct COURSE DESIGN, which consisted of questions referring to the number and length of sessions and the variety of question types.
6. A block of questions asking about the use of Backstage 2 in the course on medicine.
7. Three questions to be answered with free text asking for what students liked most / disliked about Backstage 2, and what could be done better in the future.

For all questions, a four-point Likert scale with no neutral choice was used. As for different questions different labels for the scale points were used, those will be mentioned at the respective locations in Results below to avoid confusion. In any case, the value 4 was assigned the most positive label, while 1 was assigned the most negative label.

All responses to quizzes from all participants were extracted from Backstage 2's database. Afterward, a chronological list of all responses for each user was created. For every pair of directly consecutive responses, the time difference between those was calculated. Note that this duration is *not* the duration a user took to solve a quiz but includes the time spent on reading the feedback of the previous quiz as well.

Tab. 8.1.: Overview of the participants in the course and the participants in the survey for both venues.

Course	# of participants	# of survey participants
M1	136	23
M2	103	14
Sum	239	37

From those durations, two further measures were calculated: *Sessions*, that is, a continuous sequence of responses where between two directly following responses only a short amount of time passed. For that amount, 15 minutes were chosen. Hence, sessions were determined by grouping all responses in which the duration between each pair of directly consecutive quizzes was less than 15 minutes. The second measure, *engagement duration*, that is, the overall duration a user was active in the course, was determined by adding the durations of all sessions.

When comparing more than two samples, the Kruskal-Wallis H-test with a significance threshold of $p = 0.05$ was used, as the data does not follow a normal distribution which calls for a non-parametric test (see [CF14]). Post-hoc testing was done using the Mann-Whitney U test (see [CF14]) with Bonferroni correction (see [EW07]). In this part, in case of post-hoc testing six comparisons take place (between the question types multiple choice, open answer, scale, and mark the region), and hence, the significance threshold was adapted to 0.0083 (dividing the regular significance threshold by the number of comparisons). Aggregated measures are reported as Median, hereafter abbreviated as Mdn, as it is more robust against outliers [How09], and therefore, deviation is reported as Median Absolute Deviation, hereafter abbreviated as MAD (see [RC93]).

8.2.2 Results

In the following, the results of the two aforementioned venues of the course on medicine are presented. After giving general information about the population of each course and the surveys, the students' participation in and engagement with the course, as well as the students' attitude towards the course and its components are presented.

In Table 8.1 an overview of the number of participants in the course and the survey can be seen. A user was considered a participant in the course as soon as they did at least one quiz.

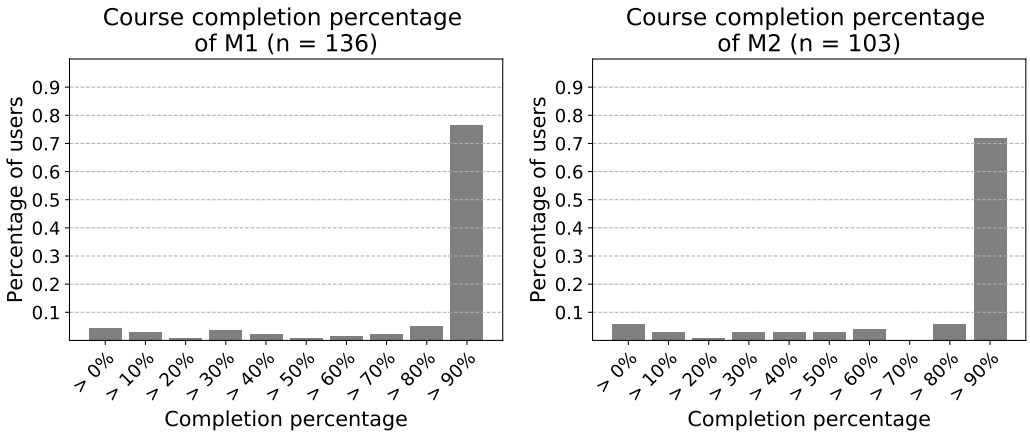


Fig. 8.2.: Percentage of users attempting the respective percentage of the course's quizzes for **M1** and **M2**.

Course Completion and Engagement Duration Figure 8.2 shows how many users completed what percentage of the course. In both **M1** and **M2**, the majority of students completed more than 90% of the course, that is, they attempted more than 90% of the available quizzes. The remainder of the students was distributed across all other percentiles with no visible trend.

Looking at the engagement duration (recall: a rough estimation of the time spent by a student working on quizzes) of those users who attempted all 90 quizzes, students in **M1** (Mdn: 108.67, MAD: 47.15) spent a bit more time in the course compared to students in **M2** (Mdn: 97.36, MAD: 37.88). In both courses, nearly the same percentage of students had an engagement duration of at least 80 minutes (70.6% in **M1**, 71.7% in **M2**).

Students' Attitude towards the Course The following paragraph presents the results of the survey and considers the responses from both courses as one.

Students were asked to rate each question type regarding its helpfulness for examination preparation (USEFULNESS), how straightforward the process of answering quizzes was (USABILITY), and how understandable the feedback given afterwards was (FEEDBACK). All answers were given on a four-point Likert scale, which was in case of USEFULNESS labelled from *not helpful at all* to *extremely helpful* and in case of USABILITY and FEEDBACK labelled from *unclear* to *clear*. The results can be seen in Table 8.2.

Regarding HELPFULNESS, students found choice quizzes to be the most helpful quizzes for examination preparation and scale quizzes the most unhelpful quizzes. The Kruskal-Wallis test indicates significant differences between the question types regarding their helpfulness for examination preparation ($p = 1.4 \cdot 10^{-8}$). While post-

Tab. 8.2.: Students' rating of each question type on the scales Helpfulness (four point Likert-scale from *not helpful at all* to *extremely helpful*), Usability, and Feedback (four point Likert-scale from *unclear* to *clear*, respectively).

	Open Answer Mdn	Scale Mdn	Mark the region Mdn	Choice Mdn
HELPFULNESS	3.0	2.0	3.0	4.0
USABILITY	4.0	4.0	3.0	4.0
FEEDBACK	3.0	3.0	3.0	4.0

hoc testing using the Mann-Whitney U test (remember, α adapted to 0.0083) revealed a few significant differences between the question types, the most striking difference was that choice quizzes were rated significantly higher regarding their usefulness for examination preparation than any other question type ($p = 1.2 \cdot 10^{-5}$ for open answer, $p = 0.0006$ for mark the region, $p = 8.2 \cdot 10^{-9}$ for scale). Regardless of that, the question types mark the region and open answer were still rated rather helpful by the majority of students. USABILITY and FEEDBACK were rated positive across all question types. While the Kruskal-Wallis test indicates significant differences for USABILITY between the question types, no post-hoc testing was done, as the results for USABILITY were positive throughout.

Summarizing those results, except for the HELPFULNESS of scale quizzes, students rated the question types across the scales HELPFULNESS, USABILITY, and FEEDBACK positively, with choice quizzes always being rated best.

The majority of students stated that they would have used the course even without the examination pictures (Mdn: 3). Students thought that the feedback showing how their peers did on a quiz helped them to better assess their knowledge (Mdn: 3), but thought that that comparison was sufficient, as they answered negatively the question which asked whether they liked to have more comparison with their peers (Mdn: 2).

A block of six questions was used to measure the construct COURSE DESIGN and included among others questions referring to the size of the sessions and the question type variety inside sessions. Students rated COURSE DESIGN positively (Mdn: 3.17, MAD: 0.25). Furthermore, the survey included six questions measuring the students' attitudes towards the course and Backstage 2. Results to these questions are shown individually in Table 8.3. Students found not only the examination images helpful for their examination preparation, but saw merit in other aspects of the system as well, such as the variety of question types and the additional value provided by Backstage 2 compared to other e-learning software used in medical education. Overall, students' attitudes towards the course design, the course itself, and Backstage 2 are positive throughout all questions.

Tab. 8.3.: Aggregated students' responses to questions measuring the attitude towards Backstage 2 and the course on medicine.

Statement	Mdn
The exam images provided on Backstage 2 were helpful for my exam preparation.	3.0
The content (besides the exam images) on Backstage 2 was helpful for my exam preparation.	3.0
Backstage 2 offered additional value (besides the exam images) not provided by any other e-learning software.	3.0
Using Backstage 2 for an exam preparation course was a good idea.	3.0
Any other e-learning software would have offered the same value as Backstage 2.	2.0
The variety of question types provided by Backstage 2 is not provided by any other e-learning software.	3.0

Finally, the survey contained three free text questions, asking what students liked or disliked about Backstage 2, and what could be done better in the future. Note that the following summary of students' answers to those questions is not a formal content analysis, but an identification of trends done by the author of this thesis.

Students mentioned specific question types (4 mentions for mark the region, 2 mentions for choice, 1 mention for open answer) as answers to the question what they liked most about Backstage 2. Furthermore, students mentioned positively the explanation texts shown after answering a quiz (2 mentions) and the variety of question types (2 mentions). Five students unspecifically mentioned pictures without making it clear whether they referred to the examination pictures or generally to the pictures used in the quizzes. In addition to those students, two students specifically mentioned examination pictures in their answer to that question.

To the question what students disliked, errors in quizzes were most frequently mentioned (5 mentions), as well as problems with mark the region quizzes (3 mentions). Indeed, the implementation of mark the region quizzes as an adapted collaborative annotation system most likely brought with it usability issues, and what is more, the library used for comparing the polygon representing the model solution with a student's polygon did not always work reliably. Moreover, students negatively mentioned the scale quizzes (3 mentions) and technical problems (3 mentions) which were fixed for **M2**.

As possible improvements, students mentioned changing the composition of the session expressed either through the suggestion of removing scale quizzes or the suggestion of adding more choice quizzes (4 mentions). Here, too, students mentioned problems with mark the region quizzes (3 mentions).

8.2.3 Discussion

Overall, the course on medicine can be considered a success: The majority of students attempted over 90% of all quizzes, students liked the course in general, the variety of question types, and the course design. Furthermore, the majority of students found that Backstage 2 offers more value than other e-learning software used in medical education. Except for scale quizzes, students liked all question types but exhibited a clear preference for choice quizzes.

The clear preference for choice quizzes is indicated by students rating the usefulness of choice quizzes for examination preparation significantly higher than the usefulness of any other question type and the explicit suggestions of adding more choice quizzes to the sessions. Students might have felt that choice quizzes prepare them best for the examination as the examination consisted exclusively of choice quizzes. Nonetheless, except for scale quizzes, all other question types were still found useful by students and were mentioned positively at least once as in the answers to the free text questions.

Generally, participation in the course was high, with over 70% of all students attempting at least 90% of all quizzes. Furthermore, the majority of students who attempted all quizzes spent an amount of time in the course (80 minutes, around 53 seconds per quiz) that suggests that those students made serious attempts at solving the quizzes as opposed to just giving a random answer to get the correct answer through the feedback shown afterward. Even though the majority of students who completed the survey stated they would have participated in the course even without the examination pictures, it is questionable whether that participation would have been of the same magnitude as the observed participation.

Another Bite-sized Learning course in another subject – Egyptology – is described in the following section. That course did not include any incentive for participation and was not necessarily only used for examination preparation but was offered throughout the term for students to refresh or acquire knowledge.

8.3 A “Catch-Up” Course on Ancient Egypt

The Institute for Egyptology and Coptology at the author’s university faces the problem that more students are studying Egyptology as their minor subject than students studying Egyptology as their major subject. Nonetheless, due to a lack of teaching staff, all those students attend the same courses, which leads to courses

where students have heterogeneous levels of knowledge in respect to the courses' contents.

To address the various levels of knowledge, a course consisting of a large number of quizzes on Ancient Egypt was created to provide students a starting point for catching up with their peers as well as an opportunity to prepare for examinations. In the following section, the different venues of that course and their differences are introduced before the results of their evaluations are discussed.

Note that this chapter refers in some places to experts of egyptology, or simply experts. Behind this are Julia Budka, Alexander Schütze, Mona Dietrich, Desiree Breineder, Eva Hemauer, and Katharina Rhymer, which helped to conceive the course, determined from which images quizzes should be generated, selected the quizzes to be included in the course, and wrote the explanation texts.

8.3.1 Venues of the Course

The course was provided to students in four terms. Both new quizzes and new question types were added between the venues of the course, and for the last venue, the course was completely overhauled. This section first introduces what all venues had in common before discussing the differences between the first three venues and the last venue.

Across all venues, the question types multiple choice, order, and locate the structure were used. Multiple choice quizzes consisted of an image and asked to either decide which structure can be seen in the image or under which king the structure shown in the image was built. Order quizzes consisted of three images and asked to order the structures shown in these images in the correct chronological order beginning with the oldest structure. An example of an order quiz can be seen in Figure 8.3: The top part shows the three images to be ordered chronologically and their assigned letters, and at the bottom, there are three blanks and three letters to be dragged into the correct blank.

The question type locate the structure, which was developed by Konrad Fischer, is constructed around a hierarchical map of Ancient Egypt on which various regions are shown as polygons. A polygon is either a target region, that is, a location of a structure, which can be submitted as answer after selecting it, or a link to a more detailed map of the region enclosed by the selected polygon. An example of the hierarchical structure can be seen in Figure 8.4: The arrows represent the process of traversing the hierarchy: A click on the region an arrow is originating from leads to

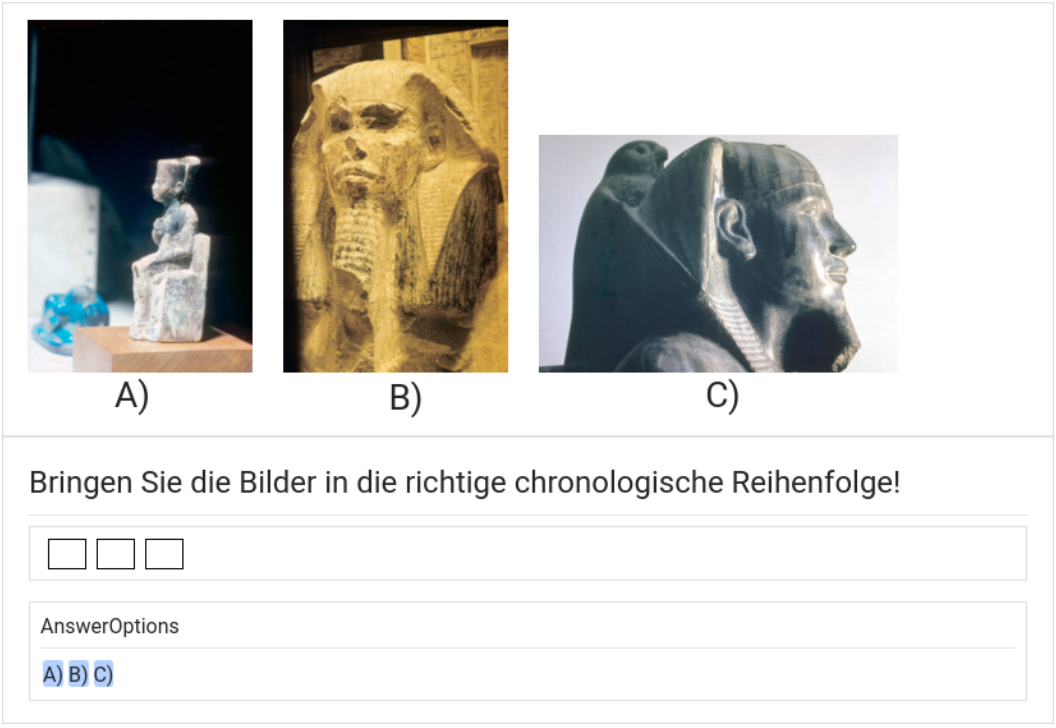


Fig. 8.3.: Example for an order quiz: The top shows the three images to be ordered chronologically and their assigned letter; the bottom part the three blanks and below that the letters which can be dragged into the correct blank (translation of the quiz question: "Arrange the images in the correct chronological order"; left image by Kurt Lange, photographers of other images unknown, all images copyright of the Institut für Ägyptologie und Koptologie of the Ludwig-Maximilians-Universität München).

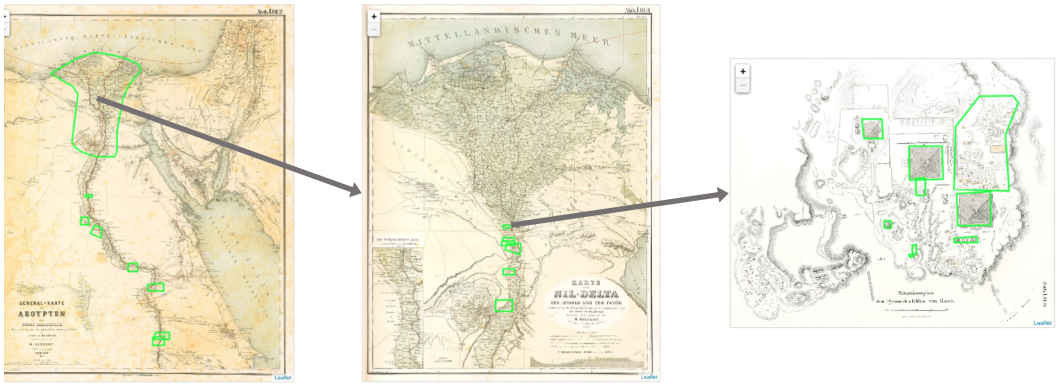


Fig. 8.4.: Overview of the hierarchical structure used by the question type locate the structure: Each arrow points to the map the click on the polygon the arrow is originating from would lead to. On all levels, there are target regions as well, which do not lead to a more detailed map, but can after selecting them be submitted as an answer (maps by Karl Richard Lepsius (1810–1884), digitalized by the Lepsius-Projekt Sachsen-Anhalt).

Tab. 8.4.: Overview of the number of quizzes and their type offered in each of the venues.

Course	Multiple choice	Order	Locate the structure	Total
EGY1	58	0	0	58
EGY2	58	8	0	66
EGY3	58	8	192	258
EGY4	58	8	192	258

the map the respective arrow is pointing to. The polygons on the last map are target regions, that is, can be selected as an answer.

As already mentioned, the number of quizzes and question types grew throughout the terms. Table 8.4 shows for each term how many quizzes of each type were available in the course. After answering a quiz, students received correctness feedback and an overview of how their peers did in that quiz. Furthermore, a text explaining what can be seen on the image(s) is shown on the feedback screen.

Creating a large number of quizzes by hand is highly time-consuming, and hence, the quizzes in this course were automatically generated from images that were tagged by experts. The following section outlines the process of automatically generating quizzes from tagged images.

Automatically Generating Quizzes Automatically generating quizzes is task of a question generation system, which is defined as a "system [that] takes, as input, a knowledge source and some specifications describing the questions to be generated" [Als+14, p. 73]. For automatically creating multiple choice questions, Huang et al. [Hua+12] specify a process encompassing three steps: Generating the question text

(referred to as *stem*), finding the correct answer (the *key*), and finally determining *distractors*, that is, the incorrect answer options shown beside the correct answer. The following section shortly outlines various approaches to automatic quiz generation before introducing the way the quizzes for the course on Ancient Egypt were generated.

Bhatia et al. [Bha+13] generate questions using Wikipedia. Their approach additionally takes into account domain specific-knowledge for determining distractors, which allows, for example, to make sure that all distractors belong to the same category as the key. Similarly, Karamis et al. [Kar+06] identify potential sentences for question generation from medical texts, transform those to questions, and choose their distractors from semantically similar words from a database of medical terms. Goto et al. [Got+10] automatically generate cloze multiple choice quizzes which are quizzes in which the question is a sentence with a word or a phrase missing (the *blank*), and the key and the distractors are possible values for the blank.

Another approach for question generation found in literature is using ontologies, which are, for example, used by Alsubait et al. [Als+14]. Their ontologies consist of several concepts and relations between those concepts, such as "a hospital is a healthcare provider" or "a teacher works in a school", as well as facts, such as "Nancy is married to David". From that ontology, various questions, such as "Give an example of a health care provider." or "Who is Mark married to?" can be generated (examples adapted to natural language and example questions taken verbatim from [Als+14, p. 75]). The authors identify distractors using the similarity of the key to potential distractors and scale the quizzes' difficulty by making the distractors more (resulting in more difficult quizzes) or less (resulting in easier quizzes) similar to the correct answer. Another approach using an ontology is the approach described by Gierl et al. [Gie+12] who use a "cognitive model" [Gie+12, p. 757] (which is basically an ontology), which models different triggers for a symptom: For each trigger, constraints on various attributes, such as days after the operation are defined. For example, a certain trigger is possible if the symptom occurs two to four days after the operation, another only after the seventh day after the operation. By modeling different triggers across various attributes, the authors can generate a large number of questions pertaining one symptom.

Brusilovsky and Sosnovksy [BS05] automatically generate "parameterized code-execution exercises" [BS05, p. 19], which are modified automatically at certain locations to generate different quizzes. Their quizzes ask for the output of a given program which are answered by students using a text field, that is, as an open answer. Similarly, Traynor and Gibson [TG05] generate code-execution quizzes where the answer is given by choosing from various answer options, that is, using multiple choice. Furthermore, they explore a novel way of creating the code of

Tab. 8.5.: Simplified records from the Mudira database.

Image number	Most exact dating	Description
1	Chephren	Taltempel des Chephren
2	Hatschepsut	unfertiger Obelisk
3	Sesostris I.	Grab des Chnumhotep II
	...	

quizzes "using a random *walk* through a very rich tree of potential programs" [TG05, p. 497]. For generating distractors, the authors adopt an approach using common misconceptions of students: The programs are mutated to model a misconception, the program is run, and the output added as a distractor. Their evaluation found that students preferred quizzes generated from templates over quizzes generated by random walks.

The approach used for generating the quizzes on Ancient Egypt described in the following is a combination of a template-based approach paired with knowledge from an ontology. The approach was conceived by Niels Heller, Elisabeth Lempa, and the author of this thesis together with the experts of egyptology; afterward, the software which creates the quizzes was implemented by Elisabeth Lempa.

Mudira[5] is a database containing a large number of images on Ancient Egypt with each of those images being tagged by an expert in various categories. Categories include original location, current location, most exact dating, and description. A simplified example of database records can be seen in Table 8.5, and is used in the following to explain how quizzes are generated automatically from those records. Note that the quiz generation process described in the following represents a simplified version and that special cases, such as missing tags, are omitted in the explanation.

Chronologically, Ancient Egypt is divided into kingdoms, which are in turn divided into dynasties, which are in turn divided into kings [Uni00]. Therefore, the column *most exact dating* is a value from one of the three levels. The column *description* represents, in most cases, what can be seen in the image. Those tags, together with an ontology that describes which rulers belong to which dynasty and which dynasties belong to which kingdom, can be used to generate a variety of quizzes.

Two templates for multiple choice quizzes were created: A template asking "From the time of which ruler is the object from?" which uses the column *most exact dating*, and another template asking "What can be seen in the image?" which uses the column *description*. Distractors are selected randomly from all possible values of the respective column. A more intelligent approach, similar to Alsubait et al.'s [Als+14]

[5] `http://mudira.gwi.uni-muenchen.de/`

approach, could take the similarity of the distractors and the correct answer into account. A possible similarity measure is the temporal distance, that is, quizzes might become more difficult when choosing distractors that are in close proximity to the key and vice versa.

Using the ontology on kingdoms, dynasties, and rulers, order quizzes can be created: Assuming Sesostris I. as chronologically earliest, and Hatshepsut as chronologically latest, an order quiz with 3, 1, 2 as correct order can be created. The basis for these quizzes is a template with the question text "Arrange the images in the correct chronological order!"

Simply generating all quizzes would lead to problems, due to a large number of quizzes ($n \cdot 2$ multiple choice quizzes and $\binom{n}{3}$ order quizzes for n being the number of records), and too difficult quizzes, because the epoch or the structure might not be evident in every image. Therefore, a human-in-the-loop approach was adapted: First, experts of egyptology selected several images from the Mudira database to create quizzes from, and second, from the quizzes generated from those images the experts chose the quizzes to be included in the course and wrote explanation texts for each of them.

For locate the structure quizzes, experts first created the link structure, that is, the polygons and which polygon on which map leads to which other map, and then for each structure to be located, the area on the corresponding map where the structure can be found. Using a program written by Konrad Fischer, the link structure and the locations of the structures were extracted from the data created by the experts, and locate the structure quizzes asking "Where is the shown structure located?" were automatically generated.

The course was provided to students for four terms. The following section details the development of the course over those four terms.

Course on Ancient Egypt Then

In its first version, the course on Ancient Egypt was structured similarly to the course on medicine: Six sessions, first only consisting of multiple choice quizzes, with order quizzes being added later on, resulting in 11 quizzes per session. Locate the structure quizzes were not added to the existing sessions but rather added as separate sessions grouped by the epoch of the structure to be located with 10 quizzes per session. In contrast to the course on medicine, students came not from the same face-to-face course, but a variety of face-to-face courses taught at the Institute for Egyptology

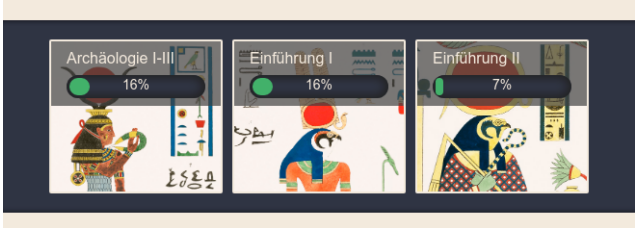


Fig. 8.5.: Overview of all presets and a user's current progress in each preset (all images by Leon Jean Joseph Dubois (1780–1846), digitalized by the New York Public Library, cropped to fit the boxes).

and Coptology. Therefore, for different students, different quizzes are relevant, and the rigid structure made it hard for students to find those quizzes that are relevant to them. Furthermore, even though three different question types were utilized, question type variety within sessions was not realized well as sessions were either containing multiple choice and order quizzes or containing only locate the structure quizzes. Hence, for the fourth venue, the course was completely overhauled to address those issues.

Course on Ancient Egypt Now

For the fourth venue, the course was completely overhauled to more closely resemble microlearning platforms, such as DuoLingo or Khan Academy. The aforementioned pre-defined sessions of 10 to 11 quizzes were replaced with adaptive sessions of 12 quizzes. Which quizzes are posed to a user in an adaptive session is dependent on a user's current learning goal and their current knowledge as assessed by the system.

Learning goals are realized through *presets*. A *preset* contains all quizzes that meet certain conditions, such as quizzes of a certain epoch and location. For identifying the quizzes included in preset, the expert tags of images were used. The view where presets can be chosen is shown in Figure 8.5: Three presets were provided to students, each preset containing the quizzes referring to the contents of certain face-to-face courses. Thus, those presets enabled students to prepare specifically for a certain face-to-face course.

During a session of 12 quizzes, an overview of the progress of the current session is given which can be seen in the top of both screenshots of Figure 8.6. Each of the 13 circles represents a quiz and the final feedback screen, respectively. A green circle represents a correctly answered question, a red circle an incorrectly answered question, and a white circle a question yet to be answered. The left screenshot of Figure 8.6 shows a multiple choice quiz, the right screenshot the correctness

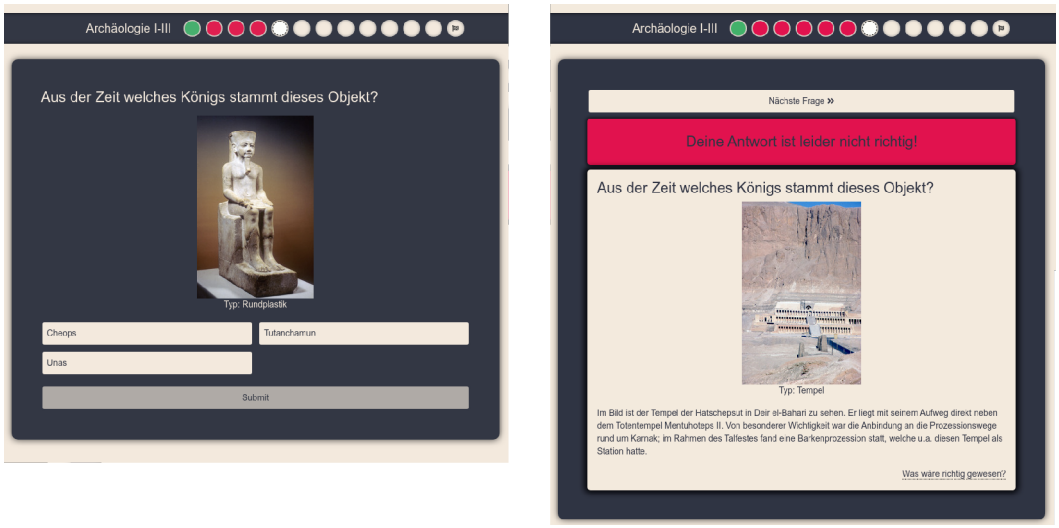


Fig. 8.6.: Screenshots of a running session: At the top of both screenshots is the progress bar which shows the progress in the current session; a green dot indicating a correctly answered question, a red dot an incorrectly answered question. The left screenshot shows a multiple choice quiz; the right screenshot the feedback view with correctness feedback and explanation text (left image copyright of the Institut für Ägyptologie und Koptologie of the Ludwig-Maximilians-Universität München, right image by Dietrich Wildung, copyright of the Staatliches Museum Ägyptischer Kunst München).

feedback and explanation text shown after answering a quiz. A click on the button at the top of this screenshot leads to the next quiz.

The next quiz and its difficulty are selected adaptively considering a user's current state and knowledge. The following section outlines the process of scaling the difficulty of the used question types and selecting appropriate quizzes.

Adaptively Selecting and Scaling Quizzes An adaptive system is a system that "provide[s] the adaption effect" [BM07, p. 3], that is, it "behave[s] differently for different users" [BM07, p. 3]. As examples for different behaviors, Brusilovsky and Millán [BM07] list that those systems can "select and prioritize the most relevant items", "provide adaptive navigation support", or "present the content adaptively" [BM07, p. 3]. In the context of adaptively selecting quizzes, that means that users should be presented those quizzes which are relevant to their current learning, as well as quizzes that are of an appropriate difficulty in respect to the their knowledge. According to Brusilovsky and Millán [BM07], a *user model* is the driver of adaptive systems, as only through a user model the adaption effect is enabled, which can be built among other from data collected from the user's interactions with a system and data provided by the user. In the following, approaches to adaptively selecting quizzes are introduced.

Brusilovsky et al.'s [Bru+04] adaptive system *QuizGuide* does not adaptively select quizzes but adaptively highlights topics in which quizzes lend themselves to be answered in respect to a student's current knowledge of the different topics. Compared to a term with no adaptive highlighting, the authors found that students had higher learning achievement, answered more quizzes, and answered a higher percentage of quizzes correctly.

Amarin et al. [Ama+09] adapt the difficulty of quizzes on a session level: After each session consisting of ten quizzes, the students' difficulty level for questions of that topic is recalculated. A similar approach is implemented by Ross et al. [Ros+18], who adapt the difficulty on session level as well. In their approach, sessions consist of ten quizzes, and if in such a session 90% of quizzes were answered correctly, the next level of difficulty for quizzes on that topic is unlocked. While the authors did not find evidence of the quizzes affecting students' learning achievement, students enjoyed the quizzes and the increasing difficulty. What these two approaches have in common is that the adaption is solely controlled by the results from the previous session. The approach of Chatzopoulou and Economides [CE10] works similar but on quiz level: During a session of 30 quizzes, a correct answer leads to the next question being of a higher difficulty level; an incorrect answer leads to the next question being of a lower difficulty (out of three levels of difficulty).

Barla et al. [Bar+10] use a three-step process to determine an appropriate quiz for a user: In a first step, they use a prerequisite graph together with an estimation of a user's knowledge in different domains to determine the areas of which quizzes may be selected from. In a second step, they select from those quizzes the most appropriate quizzes using item response theory. Finally, they use history-based heuristics to select the most appropriate quiz from those quizzes, such as quizzes that were not recently attempted by the user. Their first evaluation showed that students who used the adaptive quiz system performed better in examination questions referring to topics covered by adaptive quizzes, and their second evaluation suggested that low-performing students were those students who benefitted most from the adaptive quiz system.

Jonsdottir et al. [Jon+15] select a quiz from a pool of quizzes based on a student's grade. They determine the next quiz by drawing from a beta distribution over the quizzes ranked by difficulty (with easy quizzes being left and difficult quizzes being right). Their approach shifts the distribution further right depending on the student's grades, that is, students with better grades are more likely to get more difficult quizzes. In their system, users could request a next question as often as they liked, and users generally did so until they solved the eight previous questions correct. The authors presume that this behavior stems from the correctness of the last eight questions determining the amount of a grade bonus.

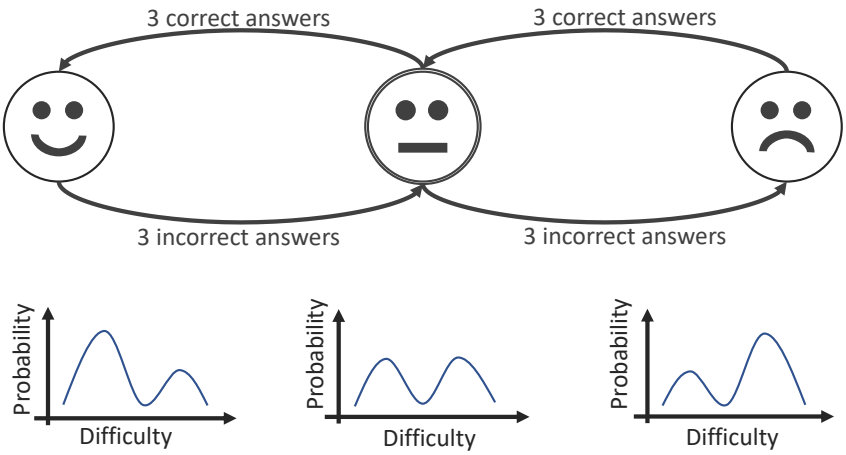


Fig. 8.7.: Simplified version of the state machine used to model the user's current state and the bimodal distribution associated with the respective states used for the adaptive selection of quizzes in the course on Ancient Egypt.

In the following, the process of the adaptive selection and difficulty scaling of question types in the course on Ancient Egypt is described. The described approach is inspired by the adaptive systems described above but includes original components as well. The concept for the adaptivity was conceived by Niels Heller, Sebastian Mader, Konrad Fischer and Korbinian Staudacher, and then implemented by Korbinian Staudacher and Konrad Fischer.

The basic idea behind the adaptive selection of quizzes is that users get easy or difficult quizzes depending on their current state. The current state is represented as a state machine which can be seen in a simplified form in the top part of Figure 8.7. Each session starts in the middle state with the number of consecutively correctly or incorrectly answered questions determining possible state changes as indicated by the edges between the states. For example, being in the initial state and answering three consecutive quizzes incorrectly would lead to changing to the state on the right. The approaches of Amarin et al. [Ama+09], Ross et al. [Ros+18], and Chatzopoulou and Economides [CE10] are similar, as they determine the difficulty of the next session or quiz (the *state*) by rules over the last few answers given by the users (the *transitions*).

Depending on the state, the difficulty of the selected quizzes changes using a bimodal distribution over the difficulty which is exemplarily shown below each state in Figure 8.7. In the initial state, it is equally likely to get an easy or difficult question. In the left state, which represents a struggling user, it is more likely to get an easy question instead of a difficult question. A distribution for determining the next quiz is used by Barla et al. [Bar+10] as well, with the difference of them using a beta distribution while the approach described here uses a bimodal distribution.

The rationale behind the state machine and the associated difficulty distributions is to model users' motivation: Not being able to solve quizzes might demotivate users, and hence, users that repeatedly fail to correctly answer quizzes are presented with easier questions to counter this aspect of demotivation. Likewise, students might become demotivated from a lack of challenge, and hence, users consecutively answering quizzes correctly are presented more difficult quizzes that challenge them.

Furthermore, as all quizzes across all question types share the same images, but the question types have inherently varying difficulties, two specific selection rules were implemented: Before an order quiz could be selected for a user, multiple choice quizzes referring to at least two of the images used in that quiz had to be answered correctly by that user. Similarly for locate the structure quizzes which are only then selected when a multiple choice quiz for the corresponding image has been answered correctly before. Those rules are similar to the prerequisite graph described by Barla et al. [Bar+10], but rather than showing which concept depends on what other concepts, these rules encode which quizzes depend on which other quizzes.

Due to the fixed number of quizzes, the described selection process can lead to a situation where no quiz with the difficulty determined by the distribution exists. Thus, for each question type, an approach for scaling down its difficulty was implemented. All those approaches aim to restrict the answer space of a quiz and by that, making it easier for users to find the correct answer. The generated multiple choice quizzes have five answer options, hence, by removing incorrect answer option by incorrect answer option, three different levels of difficulty can be generated from a single multiple choice quiz. The same approach of scaling multiple choice quizzes in difficulty is used by Papoušek and Pelanék [PP15] in their adaptive system.

In order quizzes, three structures have to be put in the correct chronological order with respect to their construction period which results in six possible answers to an order quiz. Already filling one of the blanks with the correct answer leaves two possible answers. Therefore, from one order quiz, one quiz in a lower level of difficulty can be generated. Finally, locate the structure quizzes require users to traverse a hierarchical structure of maps, where the user can – at any point – make a wrong turn and land in a part of the structure where the correct answer cannot be found. In the case where a user correctly navigates to the map on which the structure is located, there are still several answer options to chose from. Hence, there are two approaches to reduce the answer space for this question type: Letting quizzes start at a deeper (correct) level or remove answer options. However, as both approaches were not implemented in time for the last venue, locate the structure quizzes were not scaled in difficulty at all. The described mechanisms increase the

number of quizzes and the variety of difficulty which makes it more likely for a quiz with the desired difficulty to exist.

On the other hand, these mechanisms make giving students an overview of their peers' performances more difficult, as even though they might have answered the same quiz, they might have answered the quiz in different levels of difficulty. Hence, for the revamped course used in the last venue, students were no longer provided with such an overview, but were only provided correctness feedback, the correct solution, and the text explaining the correct solution.

In the following section, the results of the evaluation of the course on Ancient Egypt are introduced and discussed, but an evaluation of the adaptive selection of quizzes is omitted, as it is out of the scope of this chapter.

8.3.2 Study

As mentioned in Section 8.3.1, the course on Ancient Egypt was available in four consecutive terms (named **EGY1** to **EGY4**). The course or the quizzes newly added in the respective venue were not always available for the whole term. Especially the adaptive selection of quizzes in **EGY4** was only available for the week before the examinations. This section discusses the results from all four venues and draws further conclusions for the design of Bite-sized Learning courses.

Methods

For this evaluation, no surveys were conducted, and only data collected directly from Backstage 2's database was used. For all four venues, all quiz answers given by students and the correctness of each answer were extracted from the system. From that data, the attempted quizzes (i.e., the number of quizzes which were at least answered once) per user were calculated.

In **EGY4**, further data was available: The moment a session was started, all quizzes that were asked within a session, and the moment a session was completed, that is, a user had answered 12 quizzes. From that data, the correctness trace of a session, that is, the history of correct and incorrect answers as well as the history of question types were extracted. Furthermore, for each session, it was determined whether that session was *completed* or *abandoned*. A *completed* session is a session in which 12 quizzes were answered; otherwise the session is *abandoned*.

Tab. 8.6.: Overview of the population of each venue, the number of attempted quizzes and percentage of quizzes solved correctly at a student's first attempt.

	# of users	# of quiz attempts	# of correct on first attempt
EGY1	15	314	0.50
EGY2	12	178	0.74
EGY3	14	286	0.51
EGY4	25	501	0.40

When comparing more than two samples, the Kruskal-Wallis H-test with a significance threshold of $p = 0.05$ was used, as the data does not follow a normal distribution which calls for a non-parametric test (see [CF14]). Aggregated measures are reported as Median, hereafter abbreviated as Mdn, as it is more robust against outliers [How09], and therefore, deviation is reported as Median Absolute Deviation, hereafter abbreviated as MAD (see [RC93]).

Results

The following section presents the results of the four venues: The first paragraph discusses results across all four venues, while the last paragraph discusses **EGY4** only, as for that venue additional data was available.

Quizzes across all Venues Table 8.6 shows the number of students (that is, those students who attempted at least one quiz), the overall number of attempted quizzes, and the percentage of quizzes solved correctly by a student at their first attempt. The number of students was around 13 for the first three venues, before increasing to 26 for **EGY4**. The percentage of quizzes solved correctly at a student's first attempt hovered around 50%, with **EGY2** being an outlier caused by two students solving the majority of the course correctly in their first attempts (58 of 66 and 45 of 66, respectively). **EGY2** is not only an outlier in this respect but in the average number of attempted quizzes as well: Whereas in the other courses, that number hovered consistently around 20, (between 20.0 to 20.9), in **EGY2**, a user in average only attempted 14.8 quizzes.

Looking at the quiz attempts in more detail, Figure 8.8 shows the number of attempts on the available quizzes. Each bar represents a quiz which are ordered in the way that represents the most natural way of working through the course (units in order as they are presented). In **EGY1**, **EGY2**, and **EGY3**, where students manually selected the quizzes, the number of attempts gradually declines the further back quizzes are. Furthermore, the locate the structure quizzes, added in **EGY3**, were only attempted

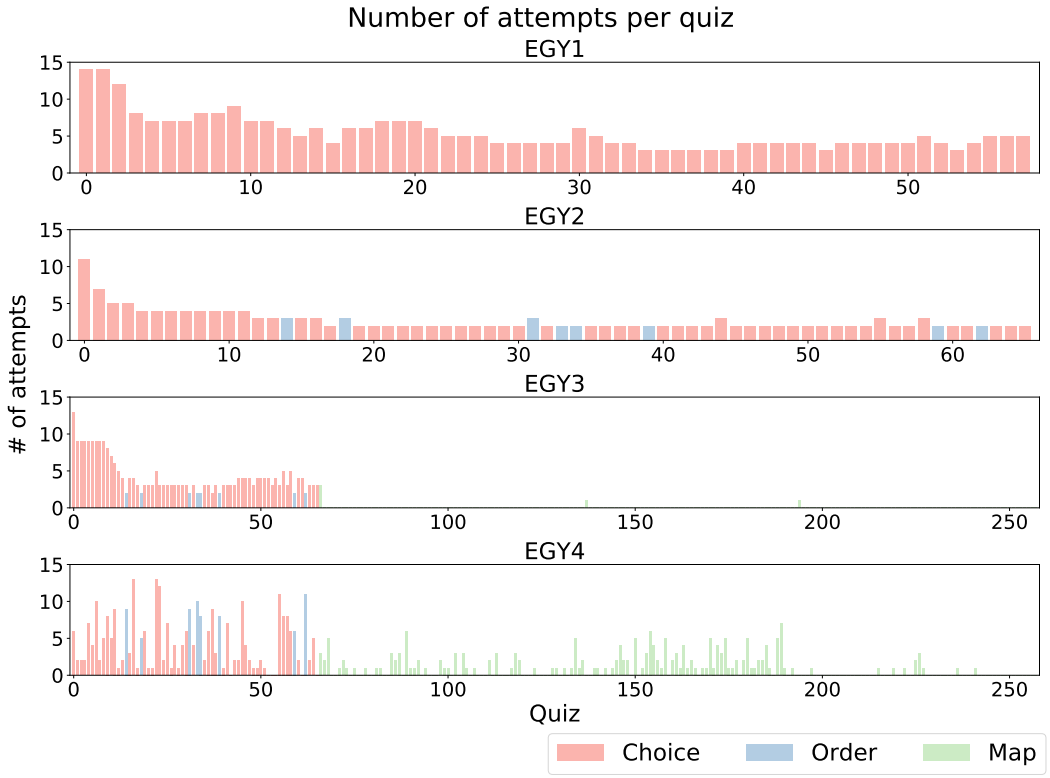


Fig. 8.8.: Number of quiz attempts per quiz for all venues. Quizzes on the x-axis are ordered by the most natural way of working through the course (units in order as they are presented).

very few times. Contrary, in **EGY4**, where quizzes were selected automatically, the attempts were more evenly distributed across all quizzes.

Regarding the number of quizzes attempted by students, Figure 8.9 shows how many quizzes were attempted by how many students. In **EGY1** and **EGY2**, there were one and two students, respectively, who attempted all quizzes. In **EGY3** and **EGY4**, there were no students who attempted all of the available quizzes, but there were two students who attempted distinctly more quizzes than their peers. In **EGY1** there was a small cluster of four students with around 40 attempted quizzes; a similar cluster of six students with around 30 attempted quizzes can be found in **EGY4**. Nonetheless, the Kruskal-Wallis test does not indicate any significant differences between the venues in regards to the number of attempted quizzes per student ($p = 0.44$).

Quiz Sessions and Adaptiveness in EGY4 Recall, that in **EGY4**, students worked through the course in sessions of 12 quizzes which were automatically selected by the system. As already mentioned, that version of course was not available for the whole term, but only starting from the week before the examinations. During that timespan, 120 sessions were started, 56 of which were completed. Figure 8.10 shows

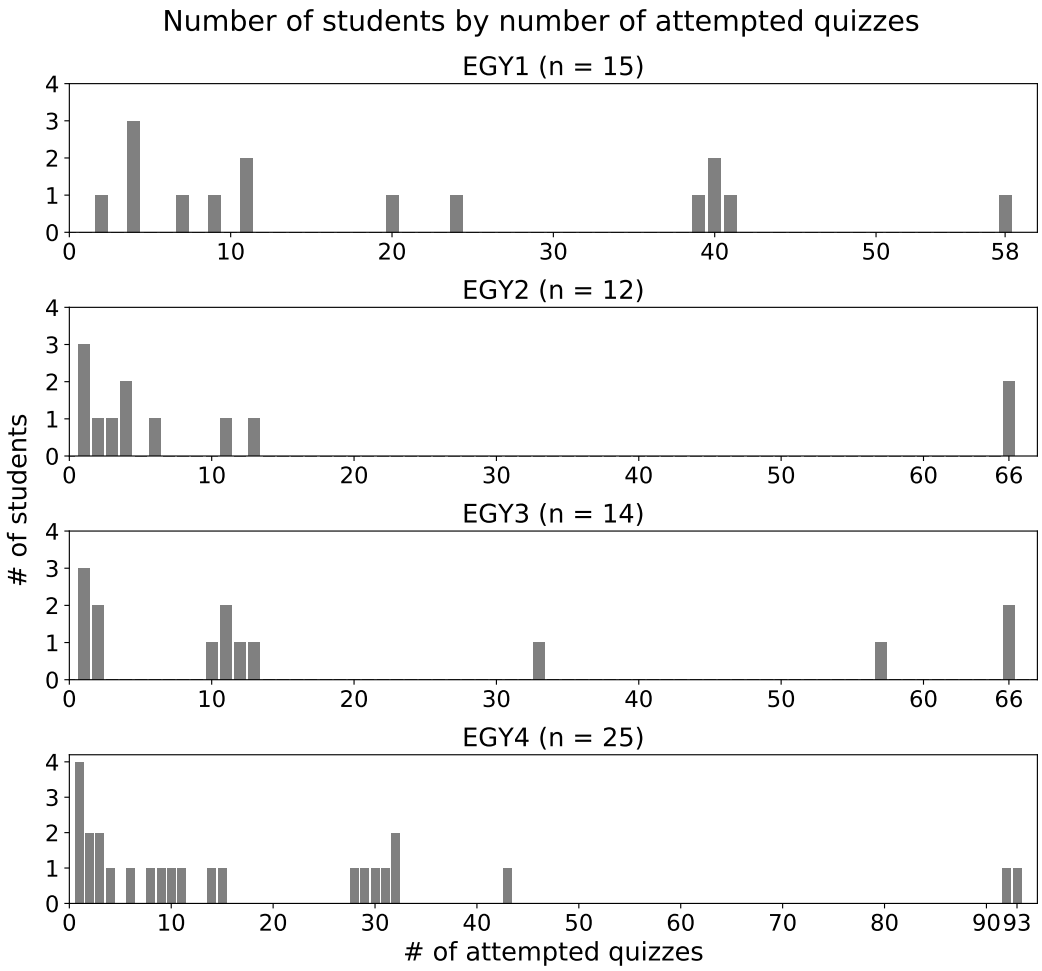


Fig. 8.9.: Number of students by number of attempted quizzes for all venues.

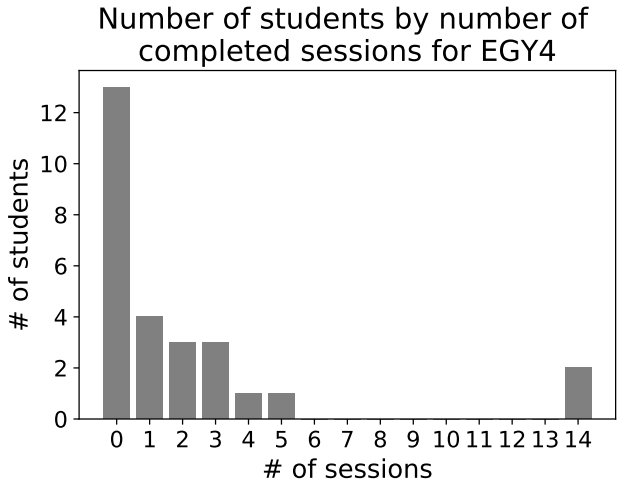


Fig. 8.10.: Number of users by number of completed sessions for **EGY4**.

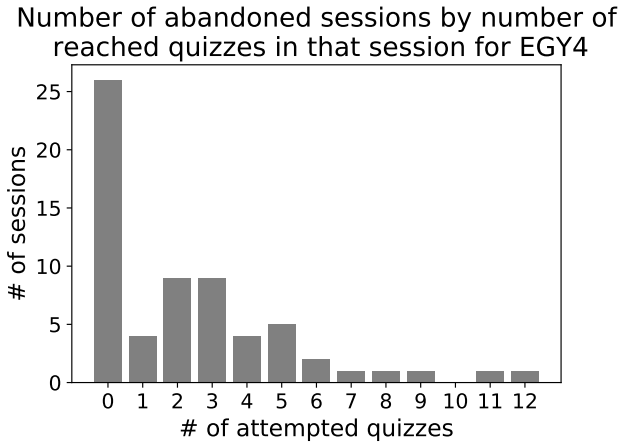


Fig. 8.11.: Number of abandoned sessions by number of quizzes answered before the session was abandoned for **EGY4**.

how many students completed how many sessions. The majority of users did not complete a single session, while the remainder of the students completed between 1 and 5 sessions. Two students completed 14 sessions. In median, a student took 4.29 minutes (MAD: 2.51) to complete a session.

Looking at more detail at abandoned sessions, which can be seen in Figure 8.11, it becomes apparent that many sessions were abandoned before a single quiz was answered, that is, the student was presented a quiz and then left Backstage 2. Generally, the majority of abandoned sessions were abandoned before half of the designated quizzes were answered.

Drilling further down into those abandoned sessions in which at least a single quiz was attempted, the correctness traces for completed sessions (green dots) and abandoned sessions (red dots) for sessions lengths (that is, the current number of

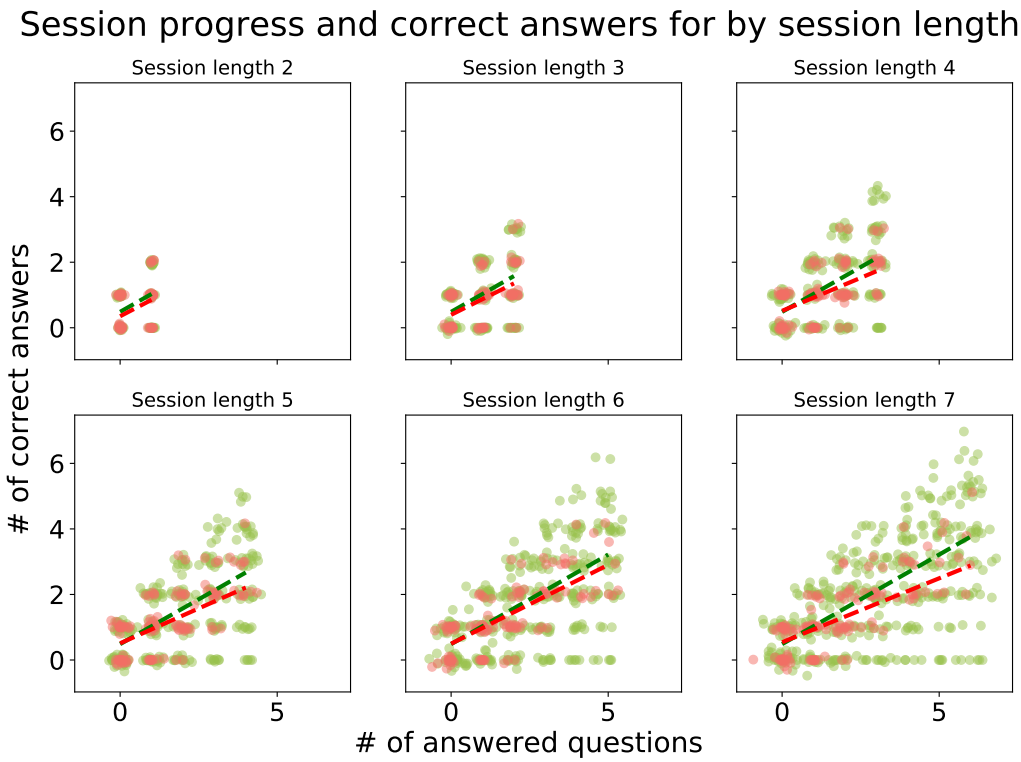


Fig. 8.12.: Correctness trace of completed (green dots) and abandoned (red dots) sessions for various session lengths with regression lines.

answered quizzes in a session) from 2 to 7 can be seen in Figure 8.12. On the x-axis, the number of answered quizzes is shown; the y-axis shows the number of quizzes answered correctly up to that point. The green and red lines are the regression lines for complete and abandoned sessions, respectively. Hence, the higher the slope of the regression line, the more quizzes have been answered correctly.

Across all session lengths, the regression line of completed sessions has always a slightly larger slope than the regression line for abandoned sessions. The higher slope of the regression lines for completed sessions suggests that students in completed sessions answered correctly slightly more quizzes. However, the difference between the regression lines is small and at no point equals a difference amounting to one or more correctly answered quizzes. Furthermore, with increasing session length, the number of abandoned sessions for that length is gradually decreasing, and for session length of 6 and 7 only 2 and 1 abandoned sessions, respectively, are added to the graphs.

Regarding the types of the selected quizzes across all sessions, the majority of those were choice quizzes (50%), followed by locate the structure (32%) and order quizzes (18%). Order quizzes were selected least often as only eight of those exist from each of which only one level of difficulty can be generated. In contrast, from one choice quiz, three levels of difficulty can be generated. While locate the structure quizzes

were not scaled in difficulty, there were much more of that type than any other type, which led to them being the second most selected question type.

The variety, that is, the periodic change of question types within a session was in median at 1.6 (MAD: 0.67) which means that around every two attempted quizzes the question type changed. However, there were six sessions for which that value is 12 which consisted exclusively of locate the structure quizzes which was caused by the sheer number of locate the structure quizzes: Once quizzes were solved correctly by a user on the highest level of difficulty, they were no longer presented to that user; a process which left after some time only locate the structure quizzes for selection.

Discussion

Four venues of the course on Ancient Egypt were presented above: Across all venues, students began attempting quizzes, but stopped after a few quizzes, only to never return to the course. Only a small number of students felt inclined to do more quizzes and return to the course. This is true for the fourth venue as well, in which students no longer selected quizzes manually, but quizzes were selected for them by the system based on their current learning goal and knowledge.

The adaptions for the fourth venue were made in the hope that the new structure and the adaptive selection of questions would motivate students to work longer with the course as quizzes were now of a more a more appropriate difficulty and more relevant to a student's current learning goal. However, the majority of students did not finish a single session and did not return to the course.

While a minority, there are a few students in all venues who attempted more quizzes than their peers, which suggests that there are a few students who saw merit in doing the quizzes. Furthermore, the adaptivity introduced in the last venue might have had a further positive effect on those students, as with answers more equally distributed across the quizzes, it is likely that they were presented more quizzes relevant to their current learning goal.

Across all venues, the only reward for doing quizzes was to get to know more about the subject matter in preparation for a face-to-face course or an examination. However, that alone seemingly was not enough to motivate students to work through the course. The possibility of the quizzes being low quality can be discarded, as quizzes were selected from the automatically generated quizzes by experts who also wrote the explanation texts. Furthermore, question types changed regularly during a

session, and hence, a lack of question type variety can likely be excluded as a reason for students abandoning the course as well.

In **EGY4**, sessions opened up the way for examining another reason for abandoning the course: Not being successful in quizzes. However, while students in completed sessions were a little bit more successful, the data (especially for abandoned sessions of higher session length) was too sparse to make a conclusive statement. Further evaluations with more data are required to make a more definite statement regarding that matter.

For future research, more data has to be collected both in the form of usage data as well as through surveys to find out why students are abandoning the course. Furthermore, this section omitted an extensive evaluation of the adaptiveness and its effects.

8.4 Wrapping up Bite-sized Learning

This chapter introduced two instances of Bite-sized Learning in different settings: A course in medicine was offered to medical students in two venues for examination preparation a week before the examination. A course on Ancient Egypt ran for four venues where in the first three, quizzes were selected manually by students, and in the last, quizzes were selected automatically by the system in respect to to a student's knowledge and learning goals. Both courses included different motivational affordances: In the course on medicine, a few images used in the quizzes were reused for questions in the examination, while in the course on Ancient Egypt the only reward was to learn more about Ancient Egypt.

While the course on medicine and the course on Ancient Egypt are not directly comparable, two conclusions can be drawn with caution from the comparison of the courses: The influence of the similarity of the quizzes in the course and the examination, as well as the influence of rewards.

Students in the course on medicine found choice quizzes, the form of questions also found in the examination, significantly more useful than any other form of quizzes. The course on Ancient Egypt provided students primarily choice quizzes as well, but, contrary to medicine, examinations in that subject tend to be in the form of short answer questions not choice questions. That fact may have promoted in students the feeling that the quizzes did not prepare them for the courses and the examinations. Using quizzes that are more similar to the types of questions asked in

examinations might be a way to motivate students to do the quizzes, but might also foster a mindset of "learning only for the examination".

Between the two instances, considerable differences in students' participation were observed, which can likely be attributed to the examination pictures included in the course on medicine. While the majority of students who participated in the survey claimed that they would have used the course even without the course containing any content included in the examination as well, it is highly questionable whether without such content a similar participation would have been observed.

There are certain limitations to the studies presented in this chapter: First, the qualitative results reported for the course on medicine are based on a small percentage of the total users who participated in the courses which might harm their validity. Furthermore, the course on medicine and the course on Ancient Egypt might not be comparable, that is, the conclusions drawn from their comparison might be invalid. Finally, the course on Ancient Egypt was not available consistently for the whole terms, which might be a reason for the reported differences between the venues.

Further research avenues in the area of Bite-sized Learning include a more detailed evaluation of the adaptive selection of quizzes and the application of the adaptive selection mechanism of quizzes in other subjects, such as the course on medicine. Furthermore, the reasons for why students abandoning such courses have to be examined to be able to develop interventions.

Bite-sized Learning closes the part of this work on technology-enhanced learning and teaching formats with a format that requires little to no interaction between teaching staff and students, as the main work of the teaching staff is the creation of quizzes. The following final part concludes this thesis: Further approaches for fostering and promoting interactivity and engagement using technology in the form of gamification and educational games are explored, and finally, the thesis is summarized and future research perspectives are outlined.

Part III

Curtain Call

The previous part demonstrated how the two main actors can be combined into technology-enhanced learning and teaching formats that break the invisible wall between lecturers and students and between students. The following last part of the thesis explores means beyond learning formats that promote interactivity and engagement both inside and outside the lecture hall in the form of educational games and gamification.

The last chapter, the final curtain call, looks back at the main actors, the learning formats, and the main results from their evaluations, before outlining perspectives for future development and research on Backstage 2. The thesis concludes with a report of a fourth (fully digital) venue of Phased Classroom Instruction in times of COVID-19 which demonstrated how technology can make difficult times less difficult.

9 Gamification and Games in Education

This chapter is based on the following articles:

- Sebastian Mader and François Bry. "Gaming the Lecture Hall: Using Social Gamification to Enhance Student Motivation and Participation". In: *The Challenges of the Digital Transformation in Education - Proceedings of the 21st International Conference on Interactive Collaborative Learning (ICL2018)*. Springer, 2018, pp. 555–566
- Sebastian Mader and François Bry. "Fun and Engagement in Lecture Halls Through Social Gamification". In: *International Journal of Engineering Pedagogy* 9.2 (2019), pp. 117–136
- Sebastian Mader, Niels Heller, and François Bry. "Adding Narrative to Gamification and Educational Games with Generic Templates". In: *Proceedings of the 18th European Conference on e-Learning (ECEL 2019)*. ACPI, 2019, pp. 360–368

Section 9.2 is based on the first two articles and adds another evaluation of the social gamification based on teams in a large class (**SG3**).
Section 9.3 is based on the last article.

When talking about the use of games in education, there are two types of games: Games built for entertainment purposes used in an educational setting and games built specifically for education either by corporations or researchers [EN06]. To varying degrees, meta-surveys on the use of games in education generally report on beneficial effects for learners [BH13; DF18; Wou+13; Ke11].

A term related to games is gamification which is "the use of game design elements in non-game contexts" [Det+11a, p. 10]. Among others, gamification is used to promote and motivate engagement with the non-game context [Kim15; SF15; Nah+14] and has been applied in a variety of contexts, such as for promoting exercise, inside work environments, and in education [Ham+14]. While several meta-surveys report – with caveats – on positive results of gamification [Ham+14; SF15; Dic+15], there is criticism as well: Robertson [Rob10] denotes gamification as "pointsification" which takes "the thing that is least essential to games and [represents] it as the core of the experience" [Rob10]. Indeed, those *least essential things*, points and badges, are among the game element most commonly used for gamification [Tod+17; SF15; Dic+15; Ham+14].

A common aspect of games and gamification in education is that they can be used to provide students with further opportunities to engage with learning activities: Games can present learning activities or subject matter in a playful manner, and gamification can provide students incentives to engage with the learning activities. Hence, to provide perspectives on promoting students' engagement beyond learning and teaching formats, this chapter explores concepts for games and gamification that promote students' engagement both inside and outside the lecture hall.

For engagement *inside* the lecture hall, a social gamification based on teams was integrated into Backstage 2's audience response system: Each student is assigned a team and contributes to their team's score by participating in and correctly answering quizzes during lecture sessions. This social gamification based on teams was evaluated with varying degrees of success in three courses.

For engagement *outside* the lecture hall, that is, engagement during the students' self-controlled study time, a generic gamification, *Reification*, and a generic educational game, *Synapses*, were devised. These approaches are generic with regard to their narrative so that a narrative that fits a course's context can be attached to them. In Reification, a learner's learning progress is visualized in the form of a landscape where learning activities are represented as objects which can either grow or decay depending on the learner's activity. Synapses uses concept maps as an additional representation of a course's contents: After each lecture session, students are tasked to organize that lecture session's contents in their ever-growing concept map. Synapses is designed as a social game: The narrative takes different turns depending on the percentage of learners who have a correct concept map. Both Reification and Synapses are mainly concepts with implementations in various stages of completion and are, subsequently, not evaluated as part of this thesis, but only illustrated by exemplary implementations.

This chapter is structured as follows: First, the literature on games in education and gamification is explored further. The following section introduces the social gamification based on teams and discusses the results of the evaluations. After that, the generic gamification Reification and the generic educational game Synapses are introduced and illustrated through exemplary implementations. In the last section of this chapter, future work and perspectives for the use of games and gamification in education are discussed.

9.1 Games and Gamification

The following section summarizes research on the use of educational games and gamification illustrated with examples.

9.1.1 Educational Games

Games and their ability to motivate and engage players make them attractive for use in education [Dic+15; Con+07]. Generally, two kinds of games are used in education: Games built for entertainment and used for education and games built specifically for education either by corporations or for research [EN06]. The first wave of educational games emerged in the 1990s but was unsuccessful, because those games simply introduced learning content into existing gameplay mechanics [Byl12]. According to de Byl [Byl12], a second wave emerged around ten years later as serious games which are no longer only used in education but in other contexts, such as health or business, as well. Indeed, the view that educational games are serious games used in the context of education is shared among many researchers (see, e.g., [Wou+13; Mun11; BH13]). For the sake of simplicity, the following chapter uses the term educational games for serious games used in education.

Even today, the characteristics de Byl [Byl12] suggested to have led to the failure of the first wave of educational games can be found in educational games: In their meta-survey on the effects of educational games, Wouters et al. [Wou+13] found that learners were not more motivated by educational games than by traditional instruction. As a possible reason for that, the authors suggest an insufficient integration of gameplay and learning content. A similar position is held by Habgood [Hab05] who argues that educational games could be more effective if gameplay and learning content were combined more naturally. As an example, he introduces a concept for a game called Zombie Division where a player defeats zombies which are labeled with numbers by selecting an attack which are labeled with numbers as well by selecting an attack which number evenly divides a zombie's number.

Several meta-surveys examined the effects of games in education: Backlund and Hendrix [BH13] conclude that games have a positive effect on learning but criticize a lack of longitudinal studies. A lack of longitudinal studies is voiced by Ke [Ke11] as well, who reports that the majority of examined studies lasted less than two hours. Regardless of that, the majority of studies examined by Ke report on significantly higher learning outcomes associated with the use games compared to traditional instruction. Wouters et al. [Wou+13] found that the use of games – especially when used over multiple sessions – led to significantly higher learning gains compared to

traditional instruction. However, their meta-survey showed that the highest learning gains were achieved when games were used in conjunction with other forms of instruction. The authors mention that generalizing from their results has to be done with caution due to the diversity of the field of games. This warning is most likely true for all of the meta-surveys described here. In the following, selected examples for educational games of different genres and used in different contexts are described.

Examples of Educational Games In Age of Computers by Natvig et al. [Nat+04], students assume the role of a time traveler and travel back to the invention of the Difference Engine (a kind of first computer). From there, students travel forwards through time visiting various periods in which important inventions in the area of computers were made. In each period, users navigate a map segmented in rooms where they have to solve exercises of various kinds. By solving exercises, users gain points which are required to travel forward to the next period. An evaluation by Natvig et al. [Nat+09] showed that students did more exercises in Age of Computers than required of them and thought that they learned more from the game than from traditional instruction. A study on learning gains from playing Age of Computers by Sindre et al. [Sin+08] revealed that students did not learn more from the game than from traditional instruction. However, the authors argue that this result does not weigh as heavy in light of students being more motivated to engage with the game than with traditional instruction.

LibraryCraft by Smith and Baker [SB11] is an educational game that teaches how to navigate and work with a library's online resources. The game is divided into several tasks embedded into a narrative. After completing a task, the narrative progresses and concludes with the player slaying a dragon after all tasks have been completed. The game was evaluated with positive results: Students thought to have learned about the library's resources by playing the game and had fun while playing the game.

Connolly et al. [Con+07] designed a game to teach students the management of software projects. In their game, each player assumes a distinct role required in the process, such as project manager or systems analyst, and each of the roles has different tasks. The systems analyst, for example, has to navigate the game world and interact with non-player characters and objects to find indications for potential requirements. After collecting and refining a list of requirements, that list is forwarded to the next player. Another game by Connolly et al. [Con+11] pertains to the area of learning foreign languages. The game is set in a world where languages are at threat of vanishing but can be saved by learners through completing tasks. By completing tasks, learners build piece by piece a contemporary Tower of Babel. An

evaluation showed that students thought to have learned something from playing the game and would play the game for longer periods.

Another game for language learning is described by Johnson [Joh07] and includes besides learning the language also the areas of learning the culture and non-verbal communication. In the game's first part, learners are taught the language, culture, and non-verbal communication and subsequently apply the knowledge in the second part of the game where they have to navigate various scripted situations in a 3D environment. The game was used to train American soldiers before being they were deployed to Iraq. The majority of players thought to have acquired a normal level of Arabic after 50 hours of playing and rated the game well.

In the area of medical education, Qin et al. [Qin+09] created a simulator in which users have to stop bleedings as well as two games unrelated to the topic that train the same psychomotor skills required to stop bleedings. In an evaluation, one group trained with the two games before having to stop a bleeding in the simulator, while the other group only trained with the simulator for the entire time. Results showed that the first group consistently and significantly outperformed the second group which suggests that training with an unrelated game that teaches some kind of skills has benefits in other situations requiring the same skills. Both groups thought that the game-based interface (e.g., an inventory, a visualization of the remaining blood, and a time limit) promoted their interest in learning blood management. Furthermore, students thought that the statistics provided to them afterward helped them to assess and improve their performance.

As an example of a non-educational game used in an educational setting, Chow et al. [Cho+11] used Deal or No Deal to teach students about expected values. An evaluation in which one group did a problem on expected values and the other group played Deal or No Deal in a group revealed that the group who played the game had higher retention.

In summary, educational games are used in a wide variety of contexts and come in various levels of sophistication, from simple gameplay where entering an answer or clicking somewhere advances the game to 3D environments who are fully navigable by the player. The results presented in this chapter suggest that, as Ke [Ke11] puts it, the question to ask is not *whether* but *how* games can be used in education. Related to educational games as described in this chapter is gamification which produces not complete games but applies elements found in those games in other contexts [Kim15].

9.1.2 Gamification

One of the various definitions of gamification proposed by Deterding et al. [Det+11a] is "the use of game design elements in non-game contexts" [Det+11a, p. 10]. A slightly earlier definition in which Deterding was involved as well includes a reason for using gamification as well, namely, "to improve user experience (UX) and user engagement" [Det+11b, p. 2425]. Indeed, increasing engagement is often mentioned as an aim and outcome of gamification (see, e.g., [SF15; Nah+14; Dic+15]).

Coming back to Deterding et al.'s [Det+11a] later definition, *game design elements* still have to be defined. For that, the authors cite Reeves and Read [RR10], who list among their ingredients for great games, "Narrative Context", "Feedback", "Reputation, Ranks, and Levels", as well as "Competition Under Rules that Are Explicit and Enforced" [RR10, p. 1f.], but argue that those elements can exist outside of games as well. As a definition for game elements, Deterding et al. [Det+11a] propose "elements that are *characteristic* to games (...), readily associated with games, and found to play a significant role in gameplay" [Det+11a, p. 12] but admit that this definition leaves much room for interpretation.

However, while Deterding et al. [Det+11a] see game elements as elements that "play a significant role in gameplay" [Det+11a, p. 12], the reality looks different: Robertson [Rob10], while referring to points and badges, claims that gamification is using "the thing that is least essential to games" [Rob10] and denotes gamification as "pointsification" [Rob10]. Indeed, points and badges and leaderboards as well are by far the most commonly used game elements in gamification [Tod+17; SF15; Dic+15] which is a kind of gamification that is criticized by Nicholson [Nic15] for being based on rewards and, therefore, only fostering extrinsic motivation.

Notwithstanding, several meta-surveys suggest that the use of gamification generally leads to positive results: The majority of studies examined by Dicheva et al. [Dic+15] report positive results on various aspects, such as engagement, attendance, and activity. However, the authors warn that gamification has to be "well designed and used correctly" [Dic+15, p. 10]. Seaborn and Fels [SF15] conclude that the studies examined by them suggest that gamification has the "potential for beneficial effects in certain contexts" [SF15, p. 29]. Of the 8 studies they examined in the area of education, five reported positive results while the other three reported mixed results. Based on that, education might be an area in which gamification might have the aforementioned "potential for beneficial effects" [SF15, p. 29]. The authors of both surveys discussed in this paragraph voice that future research should look into examining individual game elements instead of complete gamified systems to create

evidence on which game elements work and which not as well as voicing the need for more empirical research in the area. Hamari et al. [Ham+14] conclude that "gamification does work, but some caveats exist" [Ham+14, p. 3029] and that the effect is dependent on the context as well as the users of the gamified application.

Despite these positive results, gamification is, as already mentioned, criticized for being based on rewards [Nic15] or decried as "pointsification" [Rob10] or "exploitationware" [Bog11, p. 4]. The next paragraph explores criticism on gamification.

Criticism Similarly to Robertson [Rob10], Bogost [Bog11] sees points and badges (and also levels and leaderboards) not as important game elements but rather elements that are used in games to "provide structure and measure progress" [Bog11, p. 2]. Furthermore, Bogost denotes gamification as "exploitationware" [Bog11, p. 4] as users are motivated through gamification to provide some kind of real value to the provider of the gamification while the rewards they are given in return for their work have no real value (i.e., points, badges, ...).

Another problem with most gamification based on rewards in form of points and badges is that by that, according to Nicholson [Nic15], only extrinsic motivation is promoted and that "[w]hen the rewards stop, however, the behavior will likely stop" [Nic15, p. 1]. As a consequence of that, Nicholson cites Zichermann and Cunningham [ZC11] who state that to keep users doing tasks which are motivated through gamification, they "have to [be kept] (...) in that reward loop forever" [ZC11, p. 27]. Another negative aspect of gamification based on rewards is voiced by Lee and Hammer [LH11], who state that rewards might "teach students that they should only learn when provided with external rewards" [LH11, p. 4]. However, Nicholson [Nic15] argues for situations in which gamification based on rewards is appropriate as well: One of those situations is when gamification is used to motivate the learning of a skill with real-life value, that is, a skill learners see the value of after learning it. Based on that, gamification based on rewards might be suitable as a nudge to motivate users to engage in activities they might see a meaning in afterward. A similar notion is voiced by Kim [Kim15] who argues that gamification might be most successful if it is used to provide "just a little extra push to actually do the work" [Kim15, p. 34].

Independently from the aforementioned criticism, the research has suggested other negative aspects associated with gamification: Toda et al. [Tod+17] identified in a literature review four negative effects caused by gamification: "Loss of performance" [Tod+17, p. 149], where gamification had negative effects on students' learning, "Undesired behavior" [Tod+17, p. 150], where gamification failed to promote the desired learning activities, "Indifference" [Tod+17, p. 150], where gamification had

no effects, and "Declining effects" [Tod+17, p. 151], where motivation brought by gamification declined over time. Note that these authors included studies reporting positive and mixed results in their survey which means that those effects were not necessarily observed for a majority of the participants but simply reported on for the respective study. Similarly, Andrade et al. [And+16] discuss three dark sides of gamification: "Undesired Competition" [And+16, p. 179], which can have negative effects on low performing students, "Addiction and Dependence" [And+16, p. 179], with dependence potentially leading to students being unable to learn without gamification, and finally "Off-Task Behavior" [And+16, p. 178], which can happen when gamification is unrelated to the desired educational outcomes. A point similar to that last point is voiced by Callan et al. [Cal+15] who describe various exemplary gamification implementations and afterward explain how gamification is applied incorrectly in those scenarios: A constantly recurring aspect in those scenarios is a missing connection between the behavior that is motivated by the gamification and the desired outcome. According to the authors, one of the design flaws which leads to that is using "proxy behaviors" [Cal+15, p. 563] for outcomes, that is, rewarding users for actions that are presumed to have the desired outcome. The same problem is mentioned by Hung [Hun17] as well, who cite an example for an incorrect proxy behavior as well: Song and McNary [SM11] found that the number of accesses to course materials did not correlate with students' learning achievements. Hence, according to Hung [Hun17], providing gamification using such measures would not be useful, as they are not representative of the desired outcome. Another example of an incorrect proxy behavior is EcoChallenge by Ecker et al. [Eck+11] which is a system built into a car intended to promote an eco-friendly driving style. While their evaluation showed that participants were significantly more likely to engage in behaviors presumed to be eco-friendly, no significant effect on the desired outcome, the fuel consumption, could be found.

Gamification can also lead to users focussing too much on the gamification and disregarding the actual learning content which is mentioned among the undesired behaviors of Toda et al. [Tod+17] and one of the exemplary implementations of Callan et al. [Cal+15]. An example for that is the study of Halan et al. [Hal+10] who gamified a system where students chat with a simulated patient first, to learn how to do anamnesis, and second, to improve the underlying corpus (e.g., by identifying for which questions no answer or an inappropriate answer exist). They added points and leaderboards and found the gamification led to students interacting significantly more with the system. However, they also found that the interactions with the virtual patient in the group in the gamified condition focussed on getting a high score and not learning how to do anamnesis: These students were significantly less likely to greet the virtual patient and started to immediately ask questions.

Finally, for another negative aspect of gamification, Turan et al. [Tur+16] compared the cognitive load between a group of students in a gamified course and a group of students in the same course without gamification. While they observed higher engagement and higher learning gains in the gamified course, they also found that students in the gamified course had a significantly higher cognitive load. They conclude that when designing gamification, the cognitive load should be taken into account to not take away cognitive resources required for the actual learning content.

While the criticism on gamification might seem staggering at first, it mostly stems from three areas: The game elements used for gamification not being essential game elements, choosing a wrong gamification for a context, and designing gamification incorrectly. All of these issues are solvable and simply demonstrate that gamification is not a panacea for motivating users (similar voiced by [LH11; Kim15]), but can nonetheless motivate and engage users as shown by the literature discussed in the first part of this section. Hence, similar to educational games, the question might not be whether to use gamification, but how to design gamification fitting a certain context. That, however, takes much more time and effort than just slapping on points, badges, and leaderboards.

Applications GamiCAD, a gamified tutorial for the computer-aided design software AutoCAD, by Li et al. [Li+12] utilizes tasks with immediate feedback at the core of the gamification. Feedback is given in the form of a star rating and points which are based on among other completion time. Moreover, users can only progress to the next task if a certain score was achieved in the previous task. An evaluation of GamiCAD against a non-gamified version, which was a simple list of tasks with immediate feedback on correctness, showed that GamiCAD led to users completing tasks significantly faster. Moreover, users thought GamiCAD to be "more enjoyable, fun, engaging, and effective" [Li+12, p. 111].

A course on games development was set into a steampunk storyline by O'Donovan et al. [O'D+13]. Students were rewarded experience points for completing learning activities such as attending lecture sessions or completing quizzes. Experience points translated into course credit as well as into so-called SteamPoints. SteamPoints could be exchanged for various conveniences such as deadline extensions or being allowed to repeat a quiz. Furthermore, at the end of the course, those ten students with most experience points (visualized in form of a leaderboard) received a real-world reward in the form of a t-shirt. The gamification led to a significant increase in student performance and attendance during lecture sessions. Moreover, students thought that the gamification increased their engagement and supported their learning.

A world expo on a virtual island was the narrative Villagrasa et al. [Vil+14] used to gamify their course on 3D modeling. Working in groups, students were tasked to create a pavilion for that world expo from the ground up using 3D modeling software. Afterward, all pavilions were placed on the virtual island and could be visited by students using a virtual reality headset.

All of the three approaches to gamification introduced in this chapter go beyond just slapping on points, badges, and leaderboards: While GamiCAD by Li et al. [Li+12] uses points, it uses points (and in extension the star rating) to give immediate feedback in a more game-like manner and uses points for progression through the game. Similar means are found in various non-educational games, such as Angry Birds. In the course described by O'Donovan et al. [O'D+13], the authors connected rewards to things that represented a real-world value to students and used a storyline as an additional game element. Similarly to that, the gamification described by Villagrasa et al. [Vil+14] uses a narrative and introduces an explorable 3D environment (one of Reeve and Read's [RR10] ingredients for great games) as a further game element.

9.2 Gaming the Lecture Hall: Social Gamification based on Teams

The social gamification based on teams puts students into teams during lecture sessions. By participating in and correctly answering quizzes conducted during lecture sessions using the audience response system, students can contribute points to their teams' scores. Motivation is further promoted through accompanying interface components that frame the quizzes as a kind of game show: While a quiz is running, a real-time overview of team participation and incoming responses is shown to the lecture hall using a projector. In the same way, updated team standings are shown after a quiz is finished. In that sense, gamified audience response systems are, as voiced by Wang and Lieberoth [WL16], "temporarily transforming the classroom into a game show" [WL16, p. 737]. The groundwork for the team component was laid by Jacob Fürst in his (unpublished) bachelor thesis where he implemented a first version of the team component.

While this gamification uses points and leaderboards, those are only there to support competition as the main game element. While competition is seen as a game ingredient by Reeve and Read [RR10], argued for by Nicholson [Nic15] through his *engagement* dimension, and said "to motivate students through peer pressure or comparison with other students" [Mun11, p. 328], it might not be a fit for every kind of student: Andrade et al. [And+16] mention "Undesired Competition" [And+16,

p. 179] as one of their dark sides of gamification. Furthermore, there are various studies that, while generally reporting positive results, report on several students expressing negative experiences brought by the competition (see, e.g., [PL17; BK18; Tur+16]). The negative experiences most often stem from students not being able to keep up with their peers (see, e.g., [PL17; BK18]).

In more formal research, competition has been found to negatively impact on intrinsic motivation in children [Val+86], but other research suggests that the combination of competition and cooperation leads to higher task enjoyment and performance than both cooperation without competition as well as competition without cooperation [TH04]. As the proposal above combines both aspects – team members cooperate within a team and teams compete with each other – the social gamification based on teams might profit from that effect.

One prominent example for an audience response system which supports teams is Kahoot!:[1] When teams are used, students can join a team, and each team then gives a joint answer using a single device [The16]. While various studies on the use of Kahoot! generally report on positive results (see, e.g.,[PL17; BK18; Tur+16; Wan15]), it is not always evident whether those studies actually used the team component.

Bringing the discussed evidence – competition as a game element, positive effects when combining cooperation and competition, and the positive results reported for the similar system Kahoot! – together supports the suggestion that such a system could improve engagement and bring fun to lecture sessions.

The following section first introduces the implementation of teams in Backstage 2's audience response system and discusses an evaluation in a small and in a large course. Next, issues of the team component that became evident in the large course are discussed and improvements are suggested. Finally, a third evaluation with the improvements in place is reported on.

This section is a summary of already published research and omits a detailed description of the team component and certain aspects of the evaluations but adds the results and discussion of a third evaluation (**SG3**). For a more detailed discussion of the team component and the first two evaluations refer to [MB18b] and [MB19b].

Teams in Backstage 2 After a quiz has been started, an overview of all teams and each team's participation is projected in the lecture hall and updated in real-time. An example of the real-time overview of team participation can be seen in Figure 9.1

[1] `https://kahoot.com/`

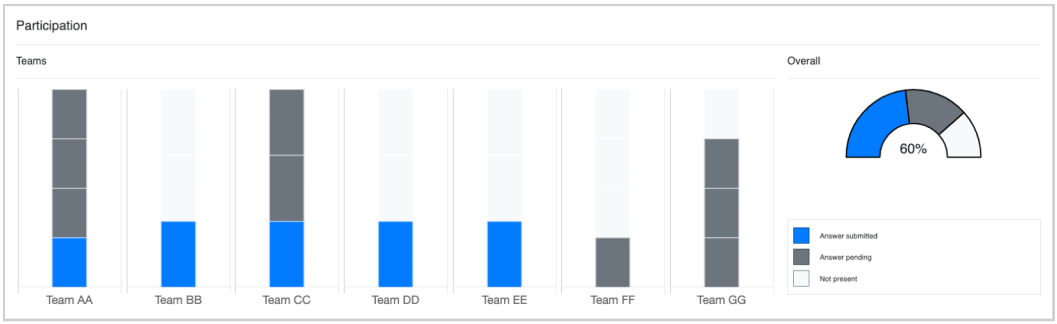


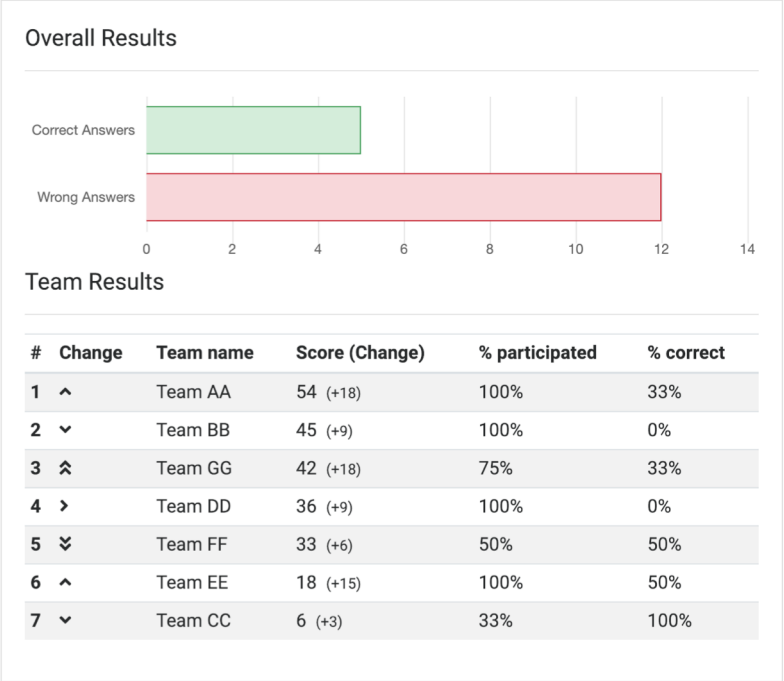
Fig. 9.1.: Screenshot of the overview projected in the lecture hall while a quiz is running. Note that this screenshot only shows the elements relevant to the gamification mechanism and omits the current quiz and the unit the quiz is attached to (adapted from [MB19b, p. 124]).

which shows the overview used in a small lecture where teams consisted of three to four members.

Each team is represented by a bar which is divided into a number of segments equivalent to the number of members in the respective team, that is, each member is represented by a segment of the bar. The color of that segment is determined by a user's state: A white segment represents a member not being present, a grey segment a member being present but not having answered yet, and a blue segment a member who has already answered. Representing each student as an individual segment provides a tangible representation of each student's contribution as opposed to, for example, a number representing the participation percentage of a team.

After a quiz has been closed by the lecturer, updated team standings are projected in the lecture hall. An example of that view can be seen in Figure 9.2. In that view, each team is represented by a row in the table and the columns represent various information about changes in standings which were caused by the just-finished quiz. For example, the column *Change* shows, similarly to standings found in sports, how the placement of the respective team changed through the points achieved in the quiz: In the example, Team AA overtook Team BB indicated by the upwards pointing wedge in Team AA's row and the downwards pointing wedge in Team BB's row. Furthermore, the table shows for each team the current score, the points acquired in the quiz, as well as the percentage of team members who participated in the quiz and the percentage of team members who gave a correct answer.

The social gamification based on teams frames a lecture session and the quizzes as a game show: Students take on the role of participants; lecturers the role of moderators. The various components are built firstly, to reinforce the feeling of being in a game show, and secondly, to support lecturers in their role as moderators: The real-time overview of team participation allows, among others, to call lagging



#	Change	Team name	Score (Change)	% participated	% correct
1	^	Team AA	54 (+18)	100%	33%
2	v	Team BB	45 (+9)	100%	0%
3	^^	Team GG	42 (+18)	75%	33%
4	>	Team DD	36 (+9)	100%	0%
5	vv	Team FF	33 (+6)	50%	50%
6	^	Team EE	18 (+15)	100%	50%
7	v	Team CC	6 (+3)	33%	100%

Fig. 9.2.: Screenshot of the updated team standings projected in the lecture hall after a quiz has been finished. Note that this screenshot only show the elements relevant to the gamification mechanism and omits the model solution and the unit the quiz is attached to (taken from [MB19b, p. 125]).

teams to action, and the team standings shown afterward provide an opportunity to comment on results and changes. As already mentioned, Wang and Lieberoth make a similar case that gamified audience response systems can make lectures feel more like game shows [WL16].

In summer term 2018, the social gamification based on teams was evaluated in two courses: The accompanying lecture sessions to a software development practical, **SG1**, which was a small class with 26 students, and a course on logics and discrete mathematics, **SG2** (already introduced as **LC1** in Chapter 5), which was a large class with around 600 students. The next section shortly outlines and discusses the evaluations of the social gamification based on teams in these two courses.

9.2.1 Initial Evaluations

Methods Among the data sources used for the evaluation of the social gamification based on teams was a survey. In **SG1**, the survey was conducted during the final lecture session; in **SG2**, the survey was conducted online after the examination which was held at the end of the course. The surveys were virtually identical in both courses and consisted of the following questions (the following list is a reproduction as found in [MB19b, p. 126f.]):

- A first group of questions referred to the students' course of study, current semester, and gender.
- A second block of questions aimed at measuring how the social gamification based on teams impacted on the motivation.
- A third block of questions aimed at measuring the engagement brought by the social gamification based on teams.
- A fourth block of questions collected self-assessments of participation.
- **SG1**: Two questions to be answered with free text which allowed students to give further feedback.
  SG2: One question to be answered with free text asked whether students felt that teams are suitable for making large lectures more engaging.

In **SG1**, answers were given on a 4-point Likert scale with no neutral choice which ranged from *strongly agree* to *strongly disagree*, while in **SG2**, answers were given on a 6-point Likert scale. To allow comparability, values from **SG1** were transformed linearly to the scale used in **SG2**. In the scale used in **SG2**, *strongly agree* was assigned the value 6 and *strongly disagree* the value 1. Note that the linear transformation leads to an issue already discussed in Chapter 7: Low values are generally overrated (*disagree* is transformed to a numerical value between *disagree* and *somewhat disagree*) and high values are generally underrated (*agree* is transformed to a numerical value between *somewhat agree* and *agree*).

In **SG2** (and later **SG3**), the questions referring to the social gamification based on teams were embedded in a larger survey which can be found in Appendix A.1. In **SG1**, a survey that only contained questions referring to the social gamification based on teams was used and can be found in Appendix A.5.

Significance was determined using the Mann-Whitney U test, as the majority of data does not follow a normal distribution which calls for a non-parametric test (see [CF14]). The significance threshold was set to $p = 0.05$. Aggregated measures are reported as Median, hereafter abbreviated as Mdn, as it is more robust against outliers [How09].

Results and Discussion Table 9.1 shows the number of students and the number of students participating in the surveys for **SG1** and **SG2**, respectively. In **SG1**, the small course, teams were created manually by the lecturer and were identical to the teams in which students worked in during the practical part of the software development practical; in **SG2**, the large course, users were randomly assigned a team upon joining the course on Backstage 2. Hence, in **SG1**, students knew their team members at the beginning of the course, while in **SG2**, students had to first

Tab. 9.1.: Overview of the population of course and survey and team sizes for **SG1** and **SG2**.

Course	# of students	# of survey participants	Team size
SG1	24	19	3 – 4
SG2	603	15 – 16	150 – 151

Tab. 9.2.: Results to the survey assessing the students' attitudes towards various aspects of the team-based social gamification in **SG1** and **SG2** (shortened versions of survey statements taken from [MB19b, p. 129]).

Statement	SG1 Mdn	SG2 Mdn
Motivating components		
Motivated by the live overview of submitted responses	4.33	4.0
Motivated by competition with other teams	4.33	2.0
Motivated by the chance to contribute to team's score	4.33	2.5
Engagement through team component		
Lecture became more engaging through the team component	4.33	3.0
Discussed answers with the team to get answer correct	4.33	2.0
Competition was fun	4.33	3.0
Participation without team component		
Would have participated without team component	4.33	5.5
Would have brought device without team component	4.33	5.0
Would prefer to solve on my own without points	2.67	3.0
Would prefer to solve on my own with points	2.67	4.5

find out who their team members were through communication outside of Backstage 2. In both courses, teams were utilized for four lecture sessions: In **SG1**, teams were used in four lecture sessions at the beginning of the term and in **SG2** for four lecture sessions in the later part of the term.

Students' attitudes towards the various aspects of the gamification can be seen in Table 9.2 which show great differences between **SG1** and **SG2**: While students in **SG2** mostly exhibited positive attitudes towards the various aspects of teams, students' attitudes in **SG2** were mostly negative.

Students in **SG1** felt motivated by the real-time overview, the competition, and the chance to contribute to their team's score, while students in **SG2** found from those elements only the real-time overview to have a somewhat motivating effect on them. Regarding the engagement brought by teams, in **SG1**, teams led to students discussing their answers with their team members and brought, in the students' opinion, fun and engagement to the lecture hall. Students in **SG2** disagreed with all those statements. While students in both courses felt that they would have

participated and brought a device to participate in quizzes even without the teams, three students in **SG1** would not have brought a device without the team component. No such students exist in **SG2**. Hence, at least in **SG1**, teams helped to increase the participation in quizzes.

The responses to the last two statements in Table 9.2 suggest that students in **SG1** liked the gamification as they preferred teams over collecting points without teams or just doing quizzes without points at all. For **SG2**, the students' answers to these statements suggest that they have a general interest in a gamified lecture hall, but that the chosen gamification (or its configuration) was not a good fit: Students would have preferred to solve the quizzes on their own with points (i.e., with some kind of gamification) rather than doing them on their own without points (i.e., with no gamification).

In summary, teams succeeded in the small course but failed in the large course. Possible reasons for that failure can be found in the students' answers to the free text question from **SG2**'s survey which asked whether teams are a suitable means to make large classes more engaging.

Nine students answered that question: Students' criticism mostly related to the configuration of the team component, such as students not being put in the same team as their friends or not knowing who their fellow team members were. One student mentioned that teams with more participating members had an inherent advantage. Indeed, participation and giving a correct answer rewarded a fixed amount of points which led to teams with more participating members automatically receiving more points. However, there were positive voices as well with two students responding that they thought teams to be suitable to bring engagement to large classes.

Even though the evidence gathered from those answers may be anecdotal, the voiced criticism was used as a starting point to rework the gamification for a second venue of the same course on logics and discrete mathematics. The next section shortly introduces the changes made and the subsequent evaluation.

9.2.2 Reworking Teams for Large Classes

The configuration of the team component in **SG2** was chosen for practical reasons: Assigning teams automatically minimizes management and communication overhead which would have occurred if students would have created and joined teams on their own. Minimizing these overheads is especially important in large classes to prevent disruptions. However, this very choice of team formation might have been

the reason for the failure of teams in **SG2**, because it led to friends not being in the same teams and students not knowing their fellow team members. Hence, the team component was reworked so that management and communication overhead are still kept to a minimum but at the same time to ensure that friends are in the same team and everybody knows their fellow team members.

As touched on previously, among the other options for team formation are letting students join teams created by lecturers on their own and letting students create teams on their own. These options were rejected for various reasons: The former option requires students to settle on a team to join and even then, students are still only aware of those students they talked to during the team formation to be in their team. The latter option might lead in large classes to a large number of teams. With an increasing number of teams, the real-time overview becomes more and more cluttered to the point of uselessness.

Hence, a third option that exploits the spatial arrangement of the lecture hall the course is held in was devised: That lecture hall is divided into three distinct wings with roughly the same amount of seats which lends itself to create a team for each wing. At the beginning of every lecture session, students were asked to join the team of the wing they sat in. That means that the duration of team affiliation changed as well: Students were no longer in the same team for the entire duration of a course but only for a single lecture session. This is necessary, as students chose their seats freely and not necessarily sit in the same wing every lecture session. Using the wings of the lecture hall for team formation addresses the main criticisms of the team component made in **SG2**: Under the reasonable assumption that friends sit next to each other, friends are in the same team. Furthermore, everybody knows their fellow team members, namely, all those students sitting in the same wing as they are.

A new version of the real-time overview of team participation was developed for the team component which can be seen in Figure 9.3: The three squares serve as a rough approximation of the layout of the lecture hall, and each student is represented by a small icon inside the square representing their wing. The icons are colored in the same way as the segments of the bars before: White for not being present, grey for being present but not having answered yet, and blue for already having answered.

Another change introduced as part of the rework was point scaling: As the updated approach no longer guarantees teams of equal size, it was important to not disadvantage smaller teams. Hence, points in a quiz are scaled so that teams with the same percentage of members giving a correct answer get the same amount of points.

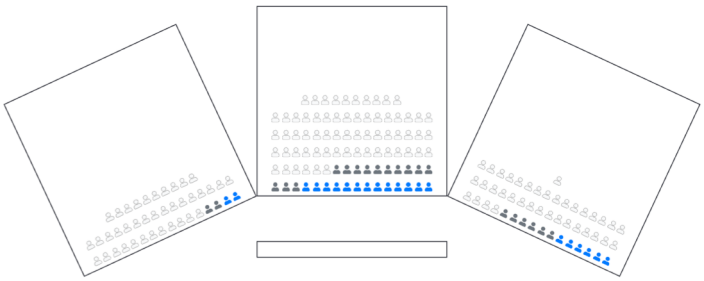


Fig. 9.3.: Revamped version of the real-time overview showing team participation (adapted from [MB19b, p. 132]).

The updated version of the team component was evaluated in another venue of the course on logics and discrete mathematics, **SG3** (already introduced as **LC2** in Chapter 5). The next section discusses the results of that evaluation.

9.2.3 Evaluating the Updated Approach

For the evaluation in **SG3**, the same survey as in **SG2** was used which made scaling unnecessary. The survey was conducted both on paper during the last lecture session as well as online to reach those students not being present during the last lecture session. Teams were used in the majority of lecture sessions, but Backstage 2 encountered technical difficulties during the first three lecture sessions which rendered it unusable for the majority of the audience.

From 609 students registered in the course, 55 participated in the survey which makes for a much better participation rate compared to **SG2**. From those 55 students, 51 to 54 students answered the questions referring to the teams.

The students' responses to the survey can be seen in Table 9.3 which includes the results of **SG2** as well for easier comparability. While the responses to all statements moved into a positive direction, only two of them flipped from a negative attitude to a positive one: Students in **SG3** rather agreed with the statement that the competition was fun and were significantly more likely to discuss with their team members to get the quiz correct ($p = 0.03$).

While results in **SG3** were still not as positive as in **SG1**, they are a clear improvement over **SG2**: Students (somewhat) liked the competition and the significant change in students' attitude towards discussion suggests that teams brought engagement in form of discussion to the lecture hall.

Wrapping up, teams were successful in the small class, failed in the first large class, but were partially successful in a subsequent large class where adaptions to approach

Tab. 9.3.: Results to the survey assessing the students' attitudes towards various aspects of the team-based social gamification in **SG2** and **SG3** (shortened versions of survey statements taken from [MB19b, p. 129]). Statements in italics indicate significant differences between the venues.

Statement	SG2 Mdn	SG3 Mdn
Engagement through team component		
Motivated by the live overview of submitted responses	4.0	5.0
Motivated by competition with other teams	2.0	3.0
Motivated by the chance to contribute to team's score	2.5	3.0
Engagement through team component		
Lecture became more engaging through the team component	3.0	3.0
Discussed answers with the team to get answer correct	2.0	4.0
Competition was fun	3.0	4.0
Engagement through team component		
Would have participated without team component	5.5	5.0
Would have brought device without team component	5.0	5.0
Would prefer to solve on my own without points	3.0	3.0
Would prefer to solve on my own with points	4.5	4.0

to team formation were made. As the biggest difference between the courses pertains the approach to team formation, it can be concluded that the way students are assigned to their teams is a critical point for the gamification: Teams seem to work better when students know who their fellow team members are and are in the same teams as their friends. One finding, however, was consistent across all courses: Students were motivated by the real-time overview of participation which suggests that simply showing students that their peers are participating motivates them to participate. Moreover, that finding seems to be independent of teams, as even in **SG2**, where teams failed, students found the real-time overview motivating.

There are limitations to the conducted evaluations which could affect the validity of the results: In **SG2**, only a small percentage of the students registered to the course participated in the survey which limits the generalizability and validity of the results. However, it must be noted, that the number of students enrolled in the courses was not representative of the number of students which took part in the evaluation of the gamification: At the time teams were introduced, around 150 to 200 students visited the weekly lecture sessions. The same applies to **SG3** where more students participated in the survey, but those are still only a minority. Furthermore, the use of the team component in **SG2** and **SG3** was not only different in respect to the described changes, but also in respect to when and for how long teams were used which might interfere with their comparability.

In future work, the gamification mechanism should be evaluated again with the same configuration and Backstage 2 working from the first lecture session without issues, as these issues could have inhibited the acceptance of Backstage 2 and deterred students from using Backstage 2 later on. Such an evaluation would provide more evidence on whether the social gamification based on teams is an appropriate means for bringing fun and engagement to lecture halls and possibly provide insights on further improvements for the team component. In any case, teams are only one possibility for gamifying audience response systems: Pohl [Poh15] envisions a gamification where students earn points depending on how fast they gave the correct answer, which conforms to the default mode of operation of Kahoot! [Kah20].

9.3 Games and Gamification outside the Lecture Hall

Educational games and gamification can bring engagement to students beyond the confinements of lecture halls as well. This section introduces a generic gamification, *Reification*, and a generic educational game, *Synapses*, that are intended to do just that – promote fun and engagement (when learning) outside of lecture sessions.

Reification visualizes a learner's learning progress in the form of a landscape where every object represents a learning activity, such as attending the weekly lecture session or turning in homework. Depending on the context Reification is used in, the landscape can take on various forms, such as an art gallery where students collect artwork, or a desert, where students collect various structures from Ancient Egypt as objects to place in their landscape.

In Synapses, students are tasked after each lecture session to organize the lecture session's contents in a concept map using concepts and associations provided by teaching staff. Moreover, Synapses attempts to address students' misconceptions by tasking students for who it is suspected that they hold a misconception to rearrange or fix the areas in their concept maps the misconception most likely stems from. Synapses is designed as a social game with a branching narrative that takes different turns depending on the percentage of students having a correct concept map.

These approaches are generic with respect to their narrative: The landscape and the objects can be freely chosen and used to tell a story that fits the context Reification is deployed in. Same for Synapses, where the storyline can be chosen to fit the context the educational game is used in. "Narrative Context" [RR10, p. 1] is, again, one of Reeve and Read's [RR10] ten ingredients for great games and can "help keep people engaged" [RR10, p. 1]. Wood [WR15] concludes in their experiment on

including a storyline into an educational game that "stories (...) are valid tools to engage learners with the learning task" [WR15, p. 326]. Further support for the use of narrative in gamification comes from Nicholson [Nic15] who argues for narrative as an element of gamification that goes beyond points and badges as part of his *exposition* dimension [Nic15]. Additionally, O'Donovan [O'D+13] argues for storyline as a gamification technique that is among the most effective ones in educational scenarios. Based on that, adding narrative might make educational games and gamification more engaging as well.

Gamification and educational games were already set in various narratives, such as a steampunk universe [O'D+13], the construction of a contemporary Tower of Babel [Con+11], a medieval story of hunting and killing a dragon [SB11], or being a time traveler who travels through the history of computers [Nat+09].

While there are instances in which the storyline is praised (see, e.g., [SB11]), there are also cases where students negatively mention that storyline and learning activities were not well connected (see, e.g., [O'D+13; Con+11]). Wouters et al. [Wou+13] even conclude in their meta-survey that including a narrative might be counterproductive but also admit that using a story "that is closely related to the learning goals might improve the effect of the narrative" [Wou+13, p. 260]. A similar notion is voiced by Callan et al. [Cal+15] who argue that when a narrative is used, it should be closely related to the context it is deployed in.

In summary, narrative might open up new layers of engagement in educational games and gamification if used correctly. The pivotal point for correct use seems to be the connection of learning activities and storyline – just slapping on a narrative on existing learning activities unrelated to the narrative might not be enough. Both Reification and Synapses are generic to such an extent to allow a variety of narratives to fit naturally. The approaches can, however, not ease the burden of actually developing the narrative, which has either be paid for [Cal+15] or created by the developers themselves which requires a healthy amount of creativity [WR15]. This section shortly introduces both concepts and illustrates every concept with an exemplary implementation. Both concepts are only partially implemented, and therefore, no evaluations were conducted.

9.3.1 Reification

Reification visualizes a learner's progress as a landscape that contains objects that represent learning activities of that learner. Reification was elaborated from an initial idea of the author of this thesis in the master thesis of Manuel Hartmann [Har18]. The following section outlines Hartmann's concept as well as new ideas.

In Reification, learners start with an empty landscape and by completing learning activities, such as attending lectures or doing homework, they receive objects that can be placed in their landscape. In that way, learners receive a visual representation of their learning progress that goes beyond achievements and progress bars in the sense that it is personal as no two landscapes look the same. Optionally, to encourage consistent activity, landscapes can slowly decay when learners show too little activity, that is, the landscape and the contained objects gradually transition to less attractive visualizations. Both components make Reification similar to social games such as FarmVille: Players plant crops on their virtual farms that decay when they are not harvested in time. However, if a player spends time, their farm gets bigger and looks better than the farms of other players spending less time.

Reification is the noun of *to reify* which means to "represent (something abstract) as a material or concrete thing" [Mer20b] and was chosen as the name of the gamification, as *something abstract*, that is, a learner's learning progress, is represented as a *material or concrete thing*, that is, the landscape and the objects in the landscape. The term *reify* has already been used in the context of gamification: Barik et al. [Bar+16] use reify to describe a similar mechanism where uninterrupted coding sessions lead to planting a tree in an ever-growing virtual forest which reifies the work done by a user. Their mechanism is inspired by the smartphone application Forest[2] which uses the same idea to stop users from using their smartphones while working. While not using the term reify, Raymer [Ray11] argues that feedback about progress is best provided graphically and describes a gamification similar to Reification where completed learning activities reward equipment for a virtual character.

As for motivation, Raymer [Ray11] argues that the motivation generated from such a system would come from tapping "into our natural instinct to collect stuff" [Ray11, p. 3]. Further motivational affordances of Reification stem from the aforementioned connection to social games, such as FarmVille. Hamari [Ham11] discusses various game mechanics found in social games and suggests that their motivational affordances stem from effects associated with loss aversion. Loss aversion is a theory stemming from behavioral economics that suggests that "losses loom larger than corresponding gains" [TK91, p. 1039]. Among the effects connected by Hamari to social games is sunk-cost fallacy, that is, "a greater tendency to continue an endeavor once an investment in money, effort, or time has been made" [AB85, p. 124]. According to Hamari, the corresponding mechanic in FarmVille is preparing the fields after which effort was made, and hence, the tendency to return increases. Another effect mentioned by Hamari is the endowment effect which suggests that "goods that are included in the individual's endowment will be more highly valued than those

[2] `https://www.forestapp.cc/`

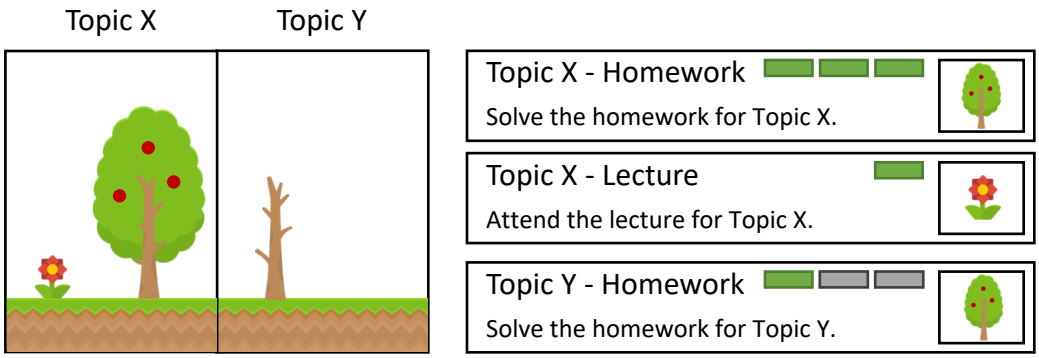


Fig. 9.4.: A landscape segmented into two topics with a completed progress and atomic task and an incomplete progress task (adapted from [Mad+19, p. 363], images taken from Kenney (`https://www.kenney.nl`)).

not held in the endowment" [Tha80, p. 44]. Hamari links that effect to the decay of objects in social games, such as the crops in FarmVille, as those have a higher value to the player just because of the fact that they already owned by the player. For a final effect, Hamari mentions the goal-gradient effect which was first observed by Hull [Hul32] and describes, as put by Kivetz et al. [Kiv+06], "that the tendency to approach a goal increases with proximity to the goal" [Kiv+06, p. 39]. In other words, the nearer the goal, the higher the motivation to complete the goal. Hamari lists progression indicators as a game mechanic that works through the goal-gradient effect. A possible example might be the crops which grow step-by-step, changing their visualization each time. How exactly these three theories relate to Reification is explained in the following part which introduces the gamification in detail.

Generic Concept

Reification visualizes learners' learning progress as landscapes filled with objects that represent various learning activities. Optionally, the landscapes can be segmented, for example, by week or topic. For the sake of simplicity, the following assumes a segmentation by topics. An example for a landscape segmented in the topics X and Y can be seen on the left side of Figure 9.4.

The right side of the same figure shows examples for tasks which are how students are assigned learning activities to complete. Each task refers to a topic and rewards an object which can be freely placed in the respective segment of the landscape. Limiting the position of the reward to a certain segment enables learners to easily identify areas which might have been given too little attention so far by searching for segments with a comparatively low number of objects. As argued by Hamari [Ham11], objects work through the sunk-cost fallacy: Completing learning objectives

to receive an object is an effort, and hence, this already expended effort might increase the tendency that a learner carries on.

There are two types of tasks: *Progress tasks* and *atomic tasks*. Progress tasks consist of more than one step. Examples of progress tasks are the first and third tasks in Figure 9.4. The number of steps and the state of a step is shown through the rectangles next to the title of the task: An already completed step is indicated by a green rectangle; an unfinished step by a grey rectangle. At the beginning of a progress task, learners are rewarded a less-developed version of the object which can already be placed in the landscape. With each completed step, the reward gradually transforms into the fully-developed version of the reward.

An example of that gradual transformation can be seen in the figure: The first task refers to topic X and all steps have been completed successfully, and hence, the reward is already a fully-grown tree carrying apples. On the other hand, the third task has only one step completed yet, and hence, the reward in the segment for topic Y is currently a less-developed tree without any leaves or apples. Completing another step would change the tree to a tree with leaves, and completing the final step to a tree with leaves and apples. Progress tasks are a kind of progress indicator and hence, according to Hamari [Ham11], utilize the goal-gradient effect: Learners are given an initial sense of progress through a less-developed object even though at that point no learning activity has taken place. Furthermore, gradual feedback on progress is given by changing to a more attractive visualization with each completed step.

Atomic tasks reward an object only after the task has been completed and consist of a single step. An example of an atomic task is the second task in Figure 9.4: After that task was completed, in this case, through visiting the lecture on Topic X, the learner was rewarded with a flower which has already been placed in the respective segment of the landscape.

For many subjects, it is undesirable that students complete learning activities for a topic once just to never return the topic. Repetition and returning to a topic are important for learning or, as phrased the other way around by Kang [Kan16], "most people know from personal experience that if one is trying to learn something well (...) a single exposure is usually inadequate for good long-term retention" [Kan16, p. 13]. A similar view is held by Polya [Pol04], who states that humans "acquire any practical skill by imitation and practice" [Pol04, p. 4]. Similar to crops in FarmVille, the objects and the landscape in Reification can decay so that learners are encouraged to return to older topics. The segment of a topic of a learner's landscape will decay if that topic is not regularly revisited by that learner, for example, by engaging in additional exercises or looking at course material on that topic. An

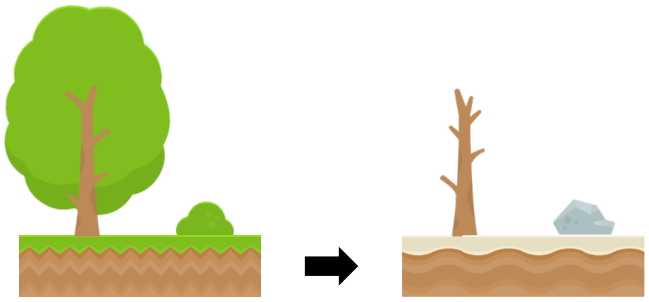


Fig. 9.5.: Example for decay in Reification: Insufficient learning activity transforms the forest into a less-attractive desert (images taken from Kenney (`https://www.kenney.nl`)).

important effect in that regard is spaced repetition which describes the effect that distributing repetitions of a learned subject matter over time generally leads to higher retention compared to repeating the subject matter (for the same amount of time) in a shorter period [Kan16]. Hence, the speed of decay and the timing when new learning activities become available have to be carefully tuned for optimal learning results.

Decay works analogously to the decay of crops in FarmVille which was connected to the endowment effect by Hamari [Ham11]: Already owned (i.e., placed) objects are overvalued by learners which might motivate them to engage in learning activities to preserve them. An example of decay can be seen in Figure 9.5 where a lush forest transitions to a dry desert which qualifies as decay under the assumption that a lush forest is more attractive to users than a dry desert.

In his master thesis, Hartmann [Har18] discusses further aspects that could strengthen the connection between learners and their landscapes: Learners could be given the possibility to create the basic layout (e.g., hills, ground, rivers, ...) of their landscape by themselves and be given the possibility to walk through their landscape with an avatar.

Reification can be given a narrative that fits the context it is used in: An art course could use an art gallery as landscape which is slowly filled with artworks, a course on geography a world map which is slowly filled with sights or landmarks, or a course on Ancient Egypt a desert which is slowly filled with structures from Ancient Egypt. The latter narrative is shortly outlined as an exemplary implementation in the next section.

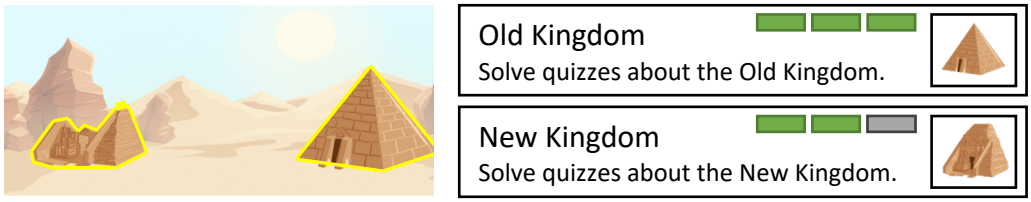


Fig. 9.6.: Concept for the implementation of Reification in the course on Ancient Egypt. On the right side, two tasks in different stages of completion can be seen. On the left side, the landscape and the objects rewarded by the tasks can be seen (adapted from [Mad+19, p. 364], landscape and structures drawn by Beatrice Sax).

Exemplary Implementation

An exemplary implementation of Reification was conceived for the Bite-sized Learning course on Ancient Egypt which was already described in Chapter 8. As narrative, the construction of structures from various epochs of Ancient Egypt was chosen. Note that even though this section is written using present tense, a working implementation of Reification for the course on Ancient Egypt does not exist. The artwork of the landscape and the structures used in the figures in this section were drawn by Beatrice Sax.

Recall, that Ancient Egypt spanned more than 5000 years which are divided into kingdoms, dynasties, and kings: A kingdom consists of a number of dynasties, which, in turn, consist of a number of kings [Uni00]. The course on Ancient Egypt consisted of a large number of quizzes from various epochs of Ancient Egypt. Each quiz was assigned the most accurate dating available, that is, either a kingdom, a dynasty, or a king.

For every kingdom (i.e., the upper-most level of the chronology), a progress task which rewards a structure in the style of the respective kingdom was created. At the beginning of each task, learners are rewarded just the foundations of the structure which can already be placed in the landscape. The structure is then completed gradually by correctly answering quizzes referring to the respective kingdom. An example of a landscape and tasks can be seen in Figure 9.6: The first task has already been completed, and hence, a completed pyramid can be seen in the landscape on the left side of the picture. The second task has only two steps completed yet, and hence, the structure is only shown in a partially completed state on the left side.

Each structure exists in three stages of completion. An example can be seen in Figure 9.7 which shows the three states of a temple in the style used during the Old Kingdom.

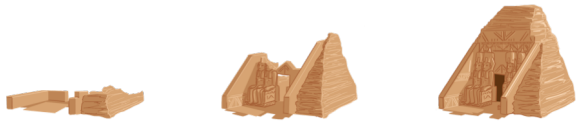


Fig. 9.7.: Different states of completion of a temple in the style used during the Old Kingdom (adapted from [Mad+19, p. 365], structures drawn by Beatrice Sax).

In the course on Ancient Egypt, no decay was used, as the intention of the course is not to get students to do quizzes regularly, but to get students to a level of knowledge which allows them to follow face-to-face courses where further learning is supposed to take place.

In summary, Reification is a generic gamification mechanism that allows attaching a narrative that fits the context it is deployed in. Furthermore, landscapes created by students are more personal than badges and achievements and that, in combination with similar motivational affordances as found in social games, might motivate students to engage in learning activities. However, Reification is just a concept with an unfinished implementation for the course on Ancient Egypt. Further research is required to consider its effects on students' motivation as well as acceptance and use among students.

9.3.2 Synapses

While Reification is a generic gamification, *Synapses* is a generic educational game in the sense that it contains elements that resemble gameplay. In Synapses, learners are tasked after a lecture session to organize the contents of the lecture session as a concept map using concepts and relationships provided by the teaching staff. Novak and Cañas [NC08] define concept maps as "graphical tools for organizing and representing knowledge" [NC08, p. 1] which are visualized as a directed graph where nodes represent concepts and edges and their labels the relationship between the connected concepts. Organizing a lecture session's contents in a concept map acts as a follow-up of the lecture session and provides students with another representation of the content. Synapses is a social game with a narrative that changes depending on the entire audience of a course: If the majority has correct concept maps, the story takes positive turns; if the majority has incorrect concept maps, the story takes negative turns. Hence, the best outcome of the story is only then achieved when the majority consistently has correct concept maps.

Furthermore, Synapses can be fully integrated into a typical face-to-face course in tertiary education that consists of lecture sessions and at-home exercises. Through

the mistakes students make in those exercises, misconceptions held by them become evident and can be addressed by interventions using a student's concept map.

Köse [Kös08] defines misconceptions as "what students themselves develop erroneously and different from scientifically accepted concepts" [Kös08, p. 283]. In their conceptual change model, Posner et al. [Pos+82] state conditions under which a misconception is abandoned for a (more) correct concept. Among those conditions is that the misconception does not longer work in a new scenario. Another condition listed by them is that there is an alternative concept available which is "intelligible", "plausible", and "fruitful" [Pos+82, p. 223]. Among approaches for triggering conceptual change, the authors mention to create situations where misconceptions are brought into conflict or to represent the subject matter differently. Another approach for addressing misconceptions are refutation texts, which consist of a description of a misconception followed by an explanation of why that is a misconception, that have been shown to be an effective way to address misconceptions [Tip10]. Novak and Cañas [NC08] mention concept maps both as a way to identify misconceptions as well as a way to address misconceptions. Indeed, concept maps have been used for identifying misconceptions held by students (see, e.g., [HP94]), as well as to quell or prevent their emergence (see, e.g., [Hsu+08; Rea+18; BS91]).

Hence, an educational game based on concept maps might improve students' learning: Organizing course's contents as a concept map provides a different representation of the subject matter and can create conflict in the students when they are not able to reflect their current understanding using the concepts and relationships provided by the teaching staff. Furthermore, tasking students to reorganize or fix areas in their concept maps where a misconception likely stems from could help to quell misconceptions.

The following section shortly outlines Synapses and describes an exemplary narrative for the course on logics and discrete mathematics already mentioned in the first part of this chapter. Note that the concept for Synapses was developed jointly with Niels Heller.

Generic Concept

There are two ideas behind Synapses: Organization of a course's contents in a concept map and the mapping of misconceptions to regions of that concept map. For the former, students are provided with concepts and labels for the edges after a lecture session which refer to the contents of that lecture session and tasked to organize these concepts as a concept map.

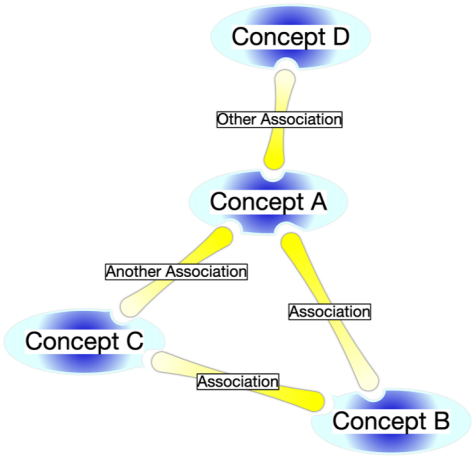


Fig. 9.8.: Display of a concept map in Synapses which is loosely inspired by how synapses in human brains actually look like. Each concept and each of its relationships represent a synapse (taken from [Mad+19, p. 365]).

An example for a concept map can be seen in Figure 9.8. Concepts are shown as blue ovals and the relationships between them are shown as yellow edges with their labels shown in the rectangles in the middle of each edge. The presentation of concept maps is (very loosely) inspired by how synapses actually look in human brains where the name of the educational game stems from. Each concept and each of its relationships represent a synapse. Points are awarded proportionally to the correctness of a student's concept map compared to a ground truth created by the teaching staff. The concept map editor shown in the figures is fully functional and was developed by Korbinian Staudacher.

Synapses is intended as a social game: While every student creates a concept map on their own, whether the majority has correct concept maps decides which turns the narrative takes: Positive turns happen if the majority has correct concept maps; negative turns if the majority has incorrect concept maps.

For the second aspect of Synapses, two things are required: Misconceptions for that subject have to be known, and these misconceptions have to be mapped to the concepts they relate to. An example from arithmetics is a misconception relating to *multiplication and division first, then addition and subtraction* which should be mapped to the concepts *Operators* and *Binding Strength*. Such a mapping makes various interventions possible: The following focusses on an intervention for face-to-face courses which comprise of lecture sessions and at-home exercises which are corrected by human tutors.

If a human tutor notices an error in a student's submission to an at-home exercise that likely stems from a misconception, the regions in the student's concept map where

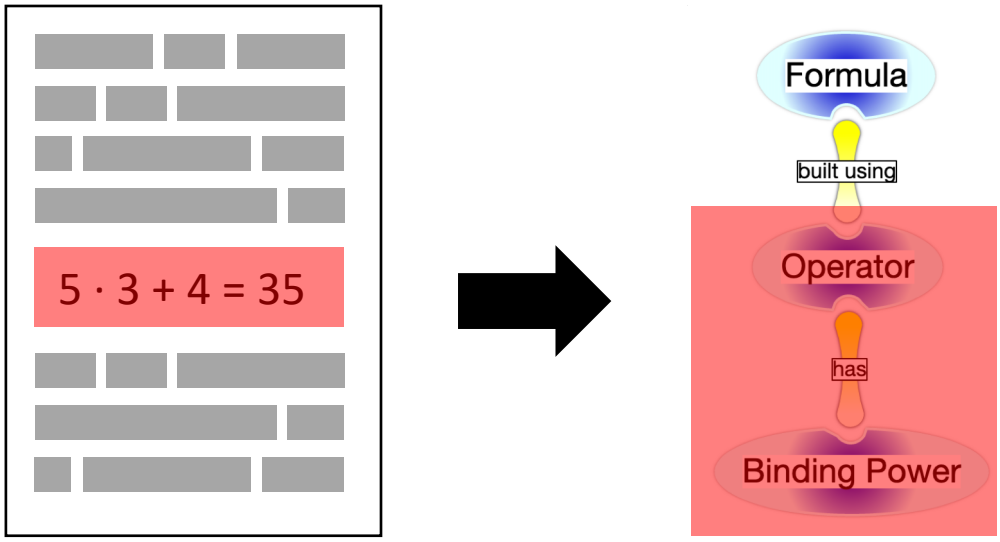


Fig. 9.9.: Process of identifying a misconception and the following intervention in Synapses: The left side shows a student's submission with a mistake likely stemming from a misconception; the right side shows the intervention which asks the students to organize the highlighted areas again (adapted from [Mad+19, p. 366]).

the related concepts are used can be identified using the aforementioned mapping. The next steps depend on the state of these areas: If these areas were already correctly organized, they are scrambled and the student is tasked to reorganize them. Likewise, if these areas are incorrectly organized, they are highlighted and the student is tasked to fix them.

An example of such an intervention using a misconception relating to *multiplication and division first, then addition and subtraction* is illustrated in Figure 9.9: The left side shows a student's submission where a mistake likely stemming from that misconception was made. Hence, the intervention, shown on the right side of the figure, highlights the areas where the concepts related to the misconception (here: Operator and Binding Power) were used and asks the student to reorganize these areas.

In his doctoral thesis, Heller [Hel20] determines the existence of a number of common errors made by students in a course on theoretical computer science. Heller found that the distribution of these errors follows a long-tail distribution, that is, few errors are done by many students, and a large number of errors are done by a few students. For Synapses, that finding implies that a mapping of misconceptions to concepts is feasible, as already a mapping of a few common misconceptions is sufficient for addressing the misconceptions of a larger number of students. Furthermore, Heller conceived and implemented a tool which allows human tutors to correct student submissions where tutors can share their corrections with their fellow tutors so that a correction for the same common error has to be written only

once. The assignment of such a correction to a student could form an entry point for triggering the aforementioned interventions on the student's concept map.

Compared to Reification, Synapses is in an earlier stage of development, hence, the next section only outlines a possible narrative but does not describe a complete implementation.

Exemplary Narrative

The narrative for Synapses described in this section is intended for the course on logics and discrete mathematics which was already discussed in the first part of this chapter. Among other topics, the course introduces logics. A real-world application of logics is the Paris Métro which has a few lines that work driverless. Parts of the correct operation of those lines were validated using the B-Method [Lec08] which is based on among others first-order logic [AP11]. That fact is mentioned more than once during the lecture sessions.

In the narrative intended for the course, Synapses is set in a fictional version of Paris where what the majority thinks is correct. As the concept maps in Synapses represent the learners' brains, they represent what the learners think. Consequently, the state of the majority's concept maps determines what is correct in that fictional world. As an example, if the majority thinks that elephants can fly, elephants would fly in that world.

Connecting the narrative with the fact that a few lines of the Paris Métro were validated using first-order logic leads to the Métro not working in the fictional world when the majority of students has an incorrect view on first-order logic, that is, has incorrect concept maps. Hence, as long as the majority has correct concept maps, the audience's fictional Paris flourishes, but when that changes, and the majority has incorrect concept maps, the Métro accumulates delays, the population grows unhappy, and the fictional Paris descends into chaos.

Both components of Synapses require an initial effort: A concept map for a course's contents has to be created which is a non-negligible effort. This concept map serves two purposes: First, as ground truth to score students' concept maps, and second, as a foundation for mapping misconceptions to concepts. To be able to create that mapping, misconceptions for a subject have to be collected first, for example from students' submissions from previous terms as done by Heller [Hel20].

Synapses is an early concept, hence, a complete implementation is required before any evaluations can be done. Regardless of that, initial evaluations should examine whether the proposed interventions help students overcome misconceptions, the organization of course's contents in a concept map supports students' learning, as well as students accept and use the educational game.

9.4 Wrapping up Gaming the Lecture Hall

This chapter was about gamification and educational games as a means for engaging students both inside and outside of lecture halls. The first part introduced a social gamification based on teams: Each student is assigned a team and contributes to their team's score by participating and correctly answering quizzes run during lecture sessions using Backstage 2's audience response system. In the second part, a generic gamification, Reification, and a generic educational game, Synapses, were introduced that are intended to be used outside of lecture sessions.

For the social gamification based on teams, evaluations in three courses were presented: The social gamification based on teams worked well in a small course but failed to varying degrees in two large courses. Despite the failure in the large courses, comparing their results with respect to the changes made to the team component suggests that when using teams in large classes, students should be in the same team as their friends and know who their fellow team members are. Another finding from the evaluations is that students across all venues felt motivated by a real-time overview of team participation which suggests that a real-time overview even without an accompanying team component could improve participation in quizzes.

Reification, a gamification mechanism introduced in the second part of this chapter, represents learners' learning progress as a landscape where objects in the landscape represent learning activities. To promote consistent learning, the landscape decays when a learner shows insufficient activity where it transitions to a less attractive representation. In Synapses, an educational game, students are tasked to organize a course's contents as a concept map after lecture sessions to provide them a different view on the contents. A student's concept map can be used as an intervention to address misconceptions held by the student: If a student makes an error which is suspected to stem from a misconception, they can be tasked to review the regions of the concept map pertaining that misconception. Both Reification and Synapses are generic in the sense that they can be used with a narrative that fits the context they are deployed in. However, they are only concepts with implementations in varying stages of completion, and hence, they have not been evaluated yet which should be done in future work.

All of the mechanisms introduced in this chapter were – if at all –, evaluated superficially, but more thorough evaluations were out of the scope of this thesis. Regardless of that, research has already established that gamification and educational games represent further avenues for bringing engagement and interactivity to lecture halls, but are means for engaging students beyond the walls of the lecture hall as well – if used correctly.

10 Summary and Perspectives

This thesis introduced the learning and teaching platform Backstage 2 and four technology-enhanced learning and teaching formats supported by Backstage 2. The learning formats aim to bring interactivity and engagement to tertiary STEM education; and evaluations have shown that they succeed in that goal. This chapter shortly summarizes the thesis and its main findings and presents future development and research perspectives.

10.1 Summary

Backstage 2 is a learning and teaching platform that aims at being a foundation for interactivity and engagement in tertiary STEM education. The main drivers of interactivity and engagement on the platform are an audience response system and a collaborative annotation system. The collaborative annotation system allows every participant of a course to annotate the learning material; the created annotations are immediately shared with all other participants who can then react to them. The audience response system allows to run quizzes of various types during lecture sessions but also for students to answer quizzes at their own pace outside the classroom. Backstage 2's audience response system extends upon today's audience response systems with respect to the supported question types which go beyond multiple choice and open answer as well as its ability to represent more complex classroom interactions through quizzes that span an arbitrary number of phases. These two components were designed with versatility in mind, so that they, together with the basic structures and features of Backstage 2, can be combined in different configurations to the four technology-enhanced learning and teaching formats.

Large Class Teaching addresses the problems of lacking interactivity and feedback in large lecture sessions: The collaborative annotation system is used as a backchannel during lecture sessions and for learning material related communication outside the classroom; the audience response system is used to bring interactivity in form of quizzes to lecture sessions and enables students to repeat the quizzes outside the classroom. Evaluations in two courses have shown that students use the platform throughout the week but mostly the days immediately preceding and following the lecture sessions with the peak being the day of the lecture session. The collaborative

annotation system was used by students both during the lecture sessions as well as outside of lecture sessions for communication. Likewise, the majority of students participated in the quizzes run during lecture sessions, but not all students logged into Backstage 2 participated in the quizzes. The option to repeat quizzes was used extensively the days before the examination. Students generally had a positive attitude towards the various aspects of the format and especially liked the quizzes.

Phased Classroom Instruction aims at scaling flipped classrooms to large audiences by supporting students and lecturers alike with technology. In this format, a mini-lecture is followed by an exercise that students work on alone or in teams while being supported by a problem- or subject-specific editor that provides them with immediate feedback and scaffolding. At the same time, lecturers are provided by technology an overview of students' progress on the exercise and suggestions whom to support. This allows lecturers to focus on supporting those students for whom the support provided by the editors is insufficient and require personal help. Phased Classroom Instruction was evaluated in three venues of a course on JavaScript with improvements made between the venues in response to the results from the previous venue's evaluation. Across all venues, students very much liked the format, the interactivity brought through the practical exercises, and preferred it vastly to a traditional lecture. Furthermore, the incremental improvements between the venues showed that the scaffolding and immediate feedback provided by the editor was indeed able to empower more students to solve exercises on their own which in turn freed up time of the lecturer to support those students for whom the support of the editor was not sufficient.

Collaborative Peer Review uses the collaborative annotation system to provide an environment for peer review that blurs the phases of traditional peer review of writings: Reviews are done in form of annotations that are immediately shared with all other participants, that is, every participant has access to all reviews while the review phase is still running. By that, misunderstandings and unclear reviews can already be addressed during the review phase, and reviewers are likely prevented from creating a same review twice. The format was evaluated across ten courses with positive results: Students preferred the approach to traditional teaching and indicated that both giving reviews and the received reviews promoted their learning. Another aspect of the format, having access to everyone's essays and reviews, was similarly well received by students. However, an examination of conversations showed that conversations often ended before issues raised in them being resolved, that is, a further reaction would have been expected from one of the communication participants which, however, very rarely happened. A reason for that might be the lack of an appropriate mechanism that notifies students of activities on their essays and reviews.

While in the previously described formats, lecturers still had a role to play – albeit one of limited importance –, in the final format explored in this thesis, *Bite-sized Learning*, the only task of lecturers is to provide students with quizzes of various types which students then can work on at their own pace. The format was used in two courses: A course on medicine consisting of 90 quizzes of a variety of types created by experts, and a course on Ancient Egypt where quizzes were automatically generated and only curated by experts. In the course on medicine, students liked the course in general, the diversity of question types, and the majority of students did nearly all of the quizzes. However, in the course on Ancient Egypt, students usually did only a few quizzes only to never return to the course. A possible explanation for that difference might lie in the rewards: In the course on medicine, students could gain an edge in the examination by working through all of the quizzes, while in the course on Ancient Egypt, there was no reward (except for obtaining knowledge).

Furthermore, besides learning and teaching formats, gamification and educational games have been explored as further means for promoting interactivity and engagement: A gamification based on teams was implemented in the audience response system where students are put into teams and contribute to their teams' scores by participating in classroom quizzes. Evaluations of the gamification suggest that it works best in smaller classes but that a real-time overview of quiz participation might be able to motivate students to participate in quizzes regardless of class size or being part of a team. Outside of Backstage 2, a generic gamification, *Reification*, and a generic educational game, *Synapses*, have been conceived. These approaches are generic with respect to their narrative, that is, they can be easily combined with a narrative that fits the context they are deployed in. In Reification, learners are rewarded objects for completing learning activities. These objects can be placed freely in a landscape that gives the learner a personalized visualization of their learning progress. Synapses has the organization of lecture content as concept maps at heart and can be embedded in a narrative that takes different turns depending on the state of the learners' concept maps. Both concepts are mainly concepts with implementations in various stages of completeness, and hence, were not evaluated.

10.2 Perspectives

There are various areas for further research in the context of Backstage 2: Existing learning formats can be extended and applied in other contexts, novel learning formats can be implemented, and the data can be used to close the feedback loop.

Learning Formats Extending existing and creating new learning formats is often associated with extending Backstage 2: Collaborative Peer Review could be improved

through a formal task assignment system which would relieve lecturers from the task of manually assigning essays to reviewers and could be used to evaluate various assignment approaches (see Heller [Hel20] who proposes to assign low-achieving students submissions from high-achieving students and vice versa). Independently from Collaborative Peer Review, a task assignment system is generally an important aspect of a learning and teaching platform and could improve the majority of the formats described in this thesis. Adding means for communication awareness would not only improve Collaborative Peer Review but Large Class Teaching as well: In Collaborative Peer Review, more communication awareness might lead to more conversations being brought to an end. In Large Class Teaching, it might promote communication outside of lecture sessions which might take place on lecture material referring to lecture sessions weeks ago where it is unlikely that other participants see it by chance. In that sense, such a mechanism would lend itself to explore and evaluate the additional value that is brought to conversations and which interface elements are suitable for promoting communication awareness. The elements proposed by Gruschke, which were shortly outlined in Chapter 7, provide a good entry point.

Phased Classroom Instruction was evaluated in a single context using a single editor: A course on JavaScript and a JavaScript editor. Another context Phased Classroom Instruction was intended to be used in are tutorial sessions: A lecturer first demonstrates how to solve an exercise of a certain problem class, followed by students working on another exercise of that problem class supported by a suitable editor. As Phased Classroom Instruction was built using the audience response system, new editors can be easily integrated but have to be implemented first. The conception and the implementation of new editors are no easy tasks, as the automatic scaffolding is heavily dependent on the exercise class, and measures that allow identifying struggling students have to be found.

While for Bite-sized Learning, an approach for an adaptive selection of quizzes has been developed, a formal evaluation and subsequent improvements to the approach were not done. A formal evaluation could be based on the correctness trace, that is, the sequence of correct and incorrect answers left by a student within a session, examine if there are similarities in traces that led to students never returning to the course, and then adapt the algorithm to prevent such traces.

Creating a Feedback Loop Through their activities on the platform, students create much data: They provide answers to quizzes, they annotate the learning material, do their homework assignments on Backstage 2 / Projects, and simply interact with the platform. This data might be valuable to lecturers, both in the present, to adapt their teaching, but in the future as well, to revise their learning material. This data,

however, is not easily accessible as it is distributed over various parts of the platform. At the beginning of the project, Backstage 2 was envisioned, together with Niels Heller, as data-driven learning and teaching platform which makes that data readily available to lecturers and by that, creates a feedback loop.

One approach would be to aggregate that data in a single view from which lecturers can gain an overview of their students what would allow them to identify at-risk students and launch appropriate interventions. In a perfect world, where lecturers have enough time that would be the approach to go. In our imperfect world, however, the solution cannot consist of giving lecturers another task to spend time on.

So, rather than burden lecturers with yet another task, the collected data can be used to make tasks which lecturers are already supposed to do easier and at the same time improve their results: Imagine a lecturer, who, while preparing lecture sessions, gets an overview of errors made in the homework submissions which then can be included into the learning material with a single click. Imagine a tutor, who, while preparing their tutorial, gets shown an overview of mistakes made by the students of their tutorial group and based on that, suggestions for exercises or topics to cover in the tutorial. Imagine a lecturer, who, while revising their course for following terms, gets an overview of the units which attracted a large number of questions and, hence, might need improvement. All of these scenarios are based on tasks that teaching staff would have to engage with either way, but their effectiveness and results might be improved through technology.

Building Backstage 2 Right Last, but not least: Backstage 2 is a prototype. To build software of that scope – even with help of contributors – required many shortcuts to be taken and strictly prioritizing what to implement. For productive use, Backstage 2 would have to be re-written from scratch, especially in regards to the configurability and combinability of the collaborative annotation system and the audience response system. Currently, nearly all configuration and combination is done via code changes, but it is conceivable that for many – or maybe even all – of that, interfaces might be possible from which persons without technical background can implement new formats or at least adapt existing formats to new contexts. However, adopting a platform of the scope of Backstage 2 for productive use is not an endeavor undertaken by a single person, but an endeavor undertaken by a team.

10.3 Closing Words

The COVID-19 pandemic posed a challenge for lecturers and students alike. Lecturers were suddenly forced to abandon face-to-face teaching for remote teaching: Among the approaches at the author's institute were slide casts (i.e., previously recorded lecture sessions), and live lecture sessions over video-conferencing software. While slide casts without any face-to-face teaching are the epitome of anonymity, isolation, and passivity, live lecture sessions over video-conferencing software reinforce these as well, as even those short moments of superficial contact with other students and lecturers are subdued by software.

In that situation, technology can be more than a simple means for video conferencing: Phased Classroom Instruction was used in conjunction with Zoom[1] and its breakout rooms to teach a fourth venue of the course on JavaScript. Breakout rooms allow lecturers to distribute participants into smaller video-conferencing rooms in which the participants can talk undisturbed from the other participants. Here, each team was assigned a breakout room, and after the mini-lecture has been held each team withdrew to their breakout room to work on the exercise. While that scenario can in no way replace a face-to-face lecture session, students were given the opportunity to talk and collaborate with their team, which brought interactivity even to the digital classroom.

Without any formal evaluation done, the results of a survey conducted during the last lecture session seem to be identical to those of the previous terms, but the correctness of teams' submissions was worse than in the last evaluated venue. There are various possible reasons for the drop in correctness: The lecturer might not have been able to help as effectively as in a real lecture hall, the collaboration between teams might have been become more difficult with usual classroom interactions, such as handing over a laptop, becoming hardly realizable, and collaboration between teams became completely impossible as each team is restricted to their breakout room. Of 84 students enrolled in the course, at least 68 were present during lecture sessions (around 81% of the enrolled students) which is a higher attendance than observed in the *physical* venues. That suggests that students valued the opportunity to interact with other students during the pandemic.

Regardless of whether Backstage 2 was used in face-to-face or remote teaching, this thesis has shown that Backstage 2 can address problems of mass classes in higher education: Students can be given a voice in large lectures, interactivity can be brought in form of quizzes, technology can make active learning usable even in

[1] `https://zoom.us/`

larger classes, students can provide formative feedback to their peers, and technology can support students' self-paced learning.

Bibliography

[AB85] Hal R Arkes and Catherine Blumer. “The psychology of sunk cost”. In: *Organizational Behavior and Human Decision Processes* 35.1 (1985), pp. 124–140 (cit. on p. 198).

[AD75] Linda Annis and J Kent Davis. “The effect of encoding and an external memory device on note taking”. In: *The Journal of Experimental Education* 44.2 (1975), pp. 44–46 (cit. on p. 16).

[Akb+10] Mostafa Akbari, Georg Böhm, and Ulrik Schroeder. “Enabling communication and feedback in mass lectures”. In: *2010 10th IEEE International Conference on Advanced Learning Technologies*. IEEE. 2010, pp. 254–258 (cit. on pp. 1, 3).

[Alq17] Emtinan Alqurashi. “Microlearning: A pedagogical approach for technology integration”. In: *The Turkish Online Journal of Educational Technology* (2017), pp. 942–947 (cit. on p. 144).

[Als+14] Tahani Alsubait, Bijan Parsia, and Uli Sattler. “Generating Multiple Choice Questions From Ontologies: Lessons Learnt”. In: *Proceedings of the 11th International Workshop on OWL: Experiences and Directions (OWLED 2014)*. Citeseer. 2014, pp. 73–84 (cit. on pp. 156–158).

[Ama+09] Ruba A Amarin, Feras Batarseh, and Issa Batarseh. “Adaptive Electronic Quizzing Method for Introductory Electrical Circuit Course.” In: *International Journal of Online Engineering* 5.3 (2009), pp. 4–7 (cit. on pp. 162, 163).

[Amr+13] Ashish Amresh, Adam R Carberry, and John Femiani. “Evaluating the effectiveness of flipped classrooms for teaching CS1”. In: *2013 IEEE Frontiers in Education Conference (FIE)*. IEEE. 2013, pp. 733–735 (cit. on p. 77).

[And+16] Fernando RH Andrade, Riichiro Mizoguchi, and Seiji Isotani. “The bright and dark sides of gamification”. In: *International Conference on Intelligent Tutoring Systems (ITS 2016)*. Springer. 2016, pp. 176–186 (cit. on pp. 184, 186).

[AP11] VS Alagar and Kasilingam Periyasamy. “The B-Method”. In: *Specification of Software Systems*. Springer, 2011, pp. 577–633 (cit. on p. 207).

[Arv14] James Arvanitakis. “Massification and the large lecture theatre: from panic to excitement”. In: *Higher Education* 67.6 (2014), pp. 735–745 (cit. on pp. 1, 2).

[AW09] Simon D Angus and Judith Watson. “Does regular online testing enhance student learning in the numerical sciences? Robust evidence from a large data set”. In: *British Journal of Educational Technology* 40.2 (2009), pp. 255–272 (cit. on p. 145).

[BA+13] Lorena Blasco-Arcas, Isabel Buil, Blanca Hernández-Ortega, and F Javier Sese. "Using clickers in class. The role of interactivity, active collaborative learning and engagement in learning performance". In: *Computers & Education* 62 (2013), pp. 102–110 (cit. on p. 3).

[BA97] J Martin Bland and Douglas G Altman. "Statistics notes: Cronbach's alpha". In: *BMJ* 314.7080 (1997), p. 572 (cit. on p. 66).

[Bab+16] Dymytro Babik, Edward F Gehringer, Jennifer Kidd, Ferry Pramudianto, and David Tinapple. "Probing the landscape: Toward a systematic taxonomy of online peer assessment systems in education". In: *Teaching & Learning Faculty Publications*. Vol. 22. 2016 (cit. on p. 121).

[Bak00] J. W. Baker. "The "Classroom Flip": Using Web course management tools to become the guide by the side". In: *Selected Papers from the 11th International Conference on College Teaching and Learning*. 2000, pp. 9–17 (cit. on p. 77).

[Ban+09] Aaron Bangor, Philip Kortum, and James Miller. "Determining what individual SUS scores mean: Adding an adjective rating scale". In: *Journal of Usability Studies* 4.3 (2009), pp. 114–123 (cit. on p. 69).

[Bar+10] Michal Barla, Mária Bieliková, Anna Bou Ezzeddinne, et al. "On the impact of adaptive test question selection for learning efficiency". In: *Computers & Education* 55.2 (2010), pp. 846–857 (cit. on pp. 162–164).

[Bar+16] Titus Barik, Emerson Murphy-Hill, and Thomas Zimmermann. "A perspective on blending programming environments and games: Beyond points, badges, and leaderboards". In: *2016 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*. IEEE. 2016, pp. 134–142 (cit. on p. 198).

[Bat10] Tony Bates. "New challenges for universities: Why they must change". In: *Changing Cultures in Higher Education*. Springer, 2010, pp. 15–25 (cit. on p. 1).

[BH13] Per Backlund and Maurice Hendrix. "Educational games-are they worth the effort? A literature survey of the effectiveness of serious games". In: *2013 5th International Conference on Games and Virtual Worlds for Serious Applications (VS-GAMES)*. IEEE. 2013, pp. 1–8 (cit. on pp. 177, 179).

[Bha+13] Arjun Singh Bhatia, Manas Kirti, and Sujan Kumar Saha. "Automatic generation of multiple choice questions using wikipedia". In: *International Conference on Pattern Recognition and Machine Intelligence*. Springer. 2013, pp. 733–738 (cit. on p. 157).

[Big11] John B Biggs. *Teaching for quality learning at university: What the student does*. McGraw-Hill Education (UK), 2011 (cit. on pp. 1, 2).

[BK18] Huseyin Bicen and Senay Kocakoyun. "Perceptions of students for gamification approach: Kahoot as a case study". In: *International Journal of Emerging Technologies in Learning (iJET)* 13.02 (2018), pp. 72–93 (cit. on p. 187).

[Bla+16] Erik Blair, Chris Maharaj, and Simone Primus. "Performance and perception in the flipped classroom". In: *Education and Information Technologies* 21.6 (2016), pp. 1465–1482 (cit. on pp. 77, 78, 80).

[Bli00] Donald A Bligh. *What's the Use of Lectures?* Jossey-Bass Publishers, 2000 (cit. on p. 2).

[Blo56] Benjamin S Bloom. *Taxonomy of educational objectives. Vol. 1: Cognitive domain*. David McKay Company, 1956 (cit. on p. 81).

[BM07] Peter Brusilovsky and Eva Millán. "User models for adaptive hypermedia and adaptive educational systems". In: *The Adaptive Web*. Springer, 2007, pp. 3–53 (cit. on p. 161).

[Bro+96] John Brooke et al. "SUS: a 'quick and dirty' usability scale". In: *Usability Evaluation In Industry* (1996), pp. 189–194 (cit. on pp. 55, 69, 245, 268).

[Bro06] Petrus MT Broersen. *Automatic autocorrelation and spectral analysis*. Springer, 2006 (cit. on p. 59).

[Bru+04] Peter Brusilovsky, Sergey Sosnovsky, and Olena Shcherbinina. "QuizGuide: Increasing the educational value of individualized self-assessment quizzes with adaptive navigation support". In: *E-Learn: World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education*. Association for the Advancement of Computing in Education (AACE). 2004, pp. 1806–1813 (cit. on p. 162).

[Bry+14] Samuel P Bryfczynski, Rebecca Brown, Josiah Hester, et al. "uRespond: iPad as interactive, personal response system". In: *Journal of Chemical Education* 91.3 (2014), pp. 357–363 (cit. on pp. 29, 30, 34, 35).

[BS05] Peter Brusilovsky and Sergey Sosnovsky. "Individualized exercises for self-assessment of programming knowledge: An evaluation of QuizPACK". In: *Journal on Educational Resources in Computing (JERIC)* 5.3 (2005) (cit. on p. 157).

[BS91] Patricia A Basili and Julie P Sanford. "Conceptual change strategies and cooperative group work in chemistry". In: *Journal of Research in Science Teaching* 28.4 (1991), pp. 293–304 (cit. on p. 204).

[BV+13] Jacob Lowell Bishop, Matthew A Verleger, et al. "The flipped classroom: A survey of the research". In: *ASEE National Conference Proceedings, Atlanta, GA*. Vol. 30. 9. 2013, pp. 1–18 (cit. on pp. 1, 77).

[Byl12] Penny de Byl. "Can digital natives level-up in a gamified curriculum". In: *Future Challenges, Sustainable Futures. ASCILITE Wellington* (2012), pp. 256–266 (cit. on p. 179).

[Cai+15] Carrie J Cai, Philip J Guo, James R Glass, and Robert C Miller. "Wait-learning: Leveraging wait time for second language education". In: *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM. 2015, pp. 3701–3710 (cit. on p. 144).

[Cal+15] Rachel C Callan, Kristina N Bauer, and Richard N Landers. "How to avoid the dark side of gamification: Ten business scenarios and their unintended consequences". In: *Gamification in Education and Business*. Springer, 2015, pp. 553–568 (cit. on pp. 184, 197).

[Cal07] Jane E Caldwell. "Clickers in the large classroom: Current research and best-practice tips". In: *CBE—Life Sciences Education* 6.1 (2007), pp. 9–20 (cit. on pp. 29, 31).

[Cam+14] Jennifer Campbell, Diane Horton, Michelle Craig, and Paul Gries. "Evaluating an inverted CS1". In: *SIGCSE '14: Proceedings of the 45th ACM technical symposium on Computer science education*. 2014, pp. 307–312 (cit. on pp. 77–80).

[Car99] Elisa Carbone. “Students behaving badly in large classes”. In: *New Directions for Teaching and Learning* 1999.77 (1999), pp. 35–43 (cit. on p. 51).

[CE10] Dimitra I Chatzopoulou and Anastasios A Economides. “Adaptive assessment of student’s knowledge in programming courses”. In: *Journal of Computer Assisted Learning* 26.4 (2010), pp. 258–269 (cit. on pp. 162, 163).

[CF14] Gregory W Corder and Dale I Foreman. *Nonparametric statistics: A step-by-step approach*. John Wiley & Sons, 2014 (cit. on pp. 57, 89, 104, 149, 166, 190).

[Cho+11] Alan F Chow, Kelly C Woodford, and Jeanne Maes. “Deal or No Deal: using games to improve student learning, retention and decision-making”. In: *International Journal of Mathematical Education in Science and Technology* 42.2 (2011), pp. 259–264 (cit. on p. 181).

[CM01] Catherine H Crouch and Eric Mazur. “Peer instruction: Ten years of experience and results”. In: *American Journal of Physics* 69.9 (2001), pp. 970–977 (cit. on p. 44).

[Cog+01] Sharon Cogdill, Tari Lin Fanderclai, Judith Kilborn, and Marian G Williams. “Backchannel: whispering in digital conversation”. In: *Proceedings of the 34th Annual Hawaii International Conference on System Sciences*. IEEE. 2001 (cit. on p. 53).

[Col+15] Lisa-Marie Collimore, Dwayne E Paré, and Steve Joordens. “SWDYT: So What Do You Think? Canadian students’ attitudes about peerScholar, an online peer-assessment tool”. In: *Learning Environments Research* 18.1 (2015), pp. 33–45 (cit. on p. 117).

[Con+07] Thomas M Connolly, Mark Stansfield, and Thomas Hainey. “An application of games-based learning within software engineering”. In: *British Journal of Educational Technology* 38.3 (2007), pp. 416–428 (cit. on pp. 179, 180).

[Con+11] Thomas M Connolly, Mark Stansfield, and Thomas Hainey. “An alternate reality game for language learning: ARGuing for multilingual motivation”. In: *Computers & Education* 57.1 (2011), pp. 1389–1415 (cit. on pp. 180, 197).

[Coo+10] Melanie M Cooper, Nathaniel Grove, Sonia M Underwood, and Michael W Klymkowsky. “Lost in Lewis structures: An investigation of student difficulties in developing representational competence”. In: *Journal of Chemical Education* 87.8 (2010), pp. 869–874 (cit. on p. 30).

[Cot+08] Sehoya H Cotner, Bruce A Fall, Susan M Wick, John D Walker, and Paul M Baepler. “Rapid feedback assessment methods: Can we improve engagement and preparation for exams in large-enrollment courses?” In: *Journal of Science Education and Technology* 17.5 (2008), pp. 437–443 (cit. on pp. 1, 2).

[Cra25a] Claude C Crawford. “Some experimental studies of the results of college note-taking”. In: *The Journal of Educational Research* 12.5 (1925), pp. 379–386 (cit. on p. 16).

[Cra25b] Claude C Crawford. “The correlation between college lecture notes and quiz papers”. In: *The Journal of Educational Research* 12.4 (1925), pp. 282–291 (cit. on p. 16).

[Cro+17] David Crosier, Peter Birch, Olga Davydovskaia, Daniela Kocanova, and Teodora Parveva. *Modernisation of Higher Education in Europe: Academic Staff – 2017*. Education, Audiovisual and Culture Executive Agency, 2017 (cit. on p. 1).

[CS07] Kwangsu Cho and Christian D Schunn. "Scaffolded writing and rewriting in the discipline: A web-based reciprocal peer review system". In: *Computers & Education* 48.3 (2007), pp. 409–426 (cit. on pp. 118, 119, 121).

[CT79] Carol A Carrier and Amy Titus. "The effects of notetaking: A review of studies." In: *Contemporary Educational Psychology* (1979), 299—314 (cit. on p. 17).

[CVM75] John F Carter and Nicholas H Van Matre. "Note taking versus note having." In: *Journal of Educational Psychology* 67.6 (1975), 900—904 (cit. on p. 17).

[DAS14] Luca De Alfaro and Michael Shavlovsky. "CrowdGrader: A tool for crowdsourcing the evaluation of homework assignments". In: *SIGCSE '14: Proceedings of the 45th ACM technical symposium on Computer science education*. 2014, pp. 415–420 (cit. on p. 119).

[Dav03] Phil Davies. "Closing the communications loop on the computerized peer-assessment of essays". In: *ALT-j* 11.1 (2003), pp. 41–54 (cit. on p. 119).

[DB04] Stephen W Draper and Margaret I Brown. "Increasing interactivity in lectures using an electronic voting system". In: *Journal of Computer Assisted Learning* 20.2 (2004), pp. 81–94 (cit. on pp. 3, 29).

[DB97] Tony Downing and Ian Brown. "Learning by cooperative publishing on the World Wide Web". In: *Active Learning* (1997), pp. 14–16 (cit. on p. 119).

[Det+11a] Sebastian Deterding, Dan Dixon, Rilla Khaled, and Lennart Nacke. "From game design elements to gamefulness: defining "Gamification"". In: *MindTrek '11: Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments*. 2011, pp. 9–15 (cit. on pp. 177, 182).

[Det+11b] Sebastian Deterding, Miguel Sicart, Lennart Nacke, Kenton O'Hara, and Dan Dixon. "Gamification: Using game-design elements in non-gaming contexts". In: *CHI EA '11: CHI '11 Extended Abstracts on Human Factors in Computing Systems*. 2011, pp. 2425–2428 (cit. on p. 182).

[DF18] Sara De Freitas. "Are games effective learning tools? A review of educational games". In: *Journal of Educational Technology & Society* 21.2 (2018), pp. 74–84 (cit. on p. 177).

[DH95] James R Davis and Daniel P Huttenlocher. "Shared annotation for cooperative learning". In: *CSCL '95: The first international conference on Computer support for collaborative learning*. 1995, pp. 84–88 (cit. on pp. 15, 18, 20).

[Dic+15] Darina Dicheva, Christo Dichev, Gennady Agre, and Galia Angelova. "Gamification in education: A systematic mapping study." In: *Journal of Educational Technology & Society* 18.3 (2015), pp. 75–88 (cit. on pp. 177, 179, 182).

[DVG72] Francis J Di Vesta and G Susan Gray. "Listening and note taking." In: *Journal of Educational Psychology* 63.1 (1972), pp. 8–14 (cit. on pp. 16, 17).

[Eck+11] Ronald Ecker, Philipp Holzer, Verena Broy, and Andreas Butz. “EcoChallenge: a race for efficiency”. In: *MobileHCI ’11: Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services*. 2011, pp. 91–94 (cit. on p. 184).

[EN06] Simon Egenfeldt-Nielsen. “Overview of research on the educational use of video games”. In: *Nordic Journal of Digital Literacy* 1.03 (2006), pp. 184–214 (cit. on pp. 177, 179).

[EW07] Alan C Elliott and Wayne A Woodward. *Statistical analysis quick reference guidebook: With SPSS examples*. Sage, 2007 (cit. on pp. 104, 149).

[Fei+12] Janet Feigenspan, Christian Kästner, Jörg Liebig, Sven Apel, and Stefan Hanenberg. “Measuring programming experience”. In: *2012 20th IEEE International Conference on Program Comprehension (ICPC)*. IEEE. 2012, pp. 73–82 (cit. on p. 88).

[Few06] Stephen Few. *Information dashboard design: The effective visual communication of data*. O’Reilly Media, Inc., 2006 (cit. on p. 14).

[FH73] Judith L Fisher and Mary B Harris. “Effect of note taking and review on recall.” In: *Journal of Educational Psychology* 65.3 (1973), 321—325 (cit. on p. 17).

[Fre+14] Scott Freeman, Sarah L Eddy, Miles McDonough, et al. “Active learning increases student performance in science, engineering, and mathematics”. In: *Proceedings of the National Academy of Sciences* 111.23 (2014), pp. 8410–8415 (cit. on p. 1).

[Fre86] Peter J Frederick. “The lively lecture – 8 variations”. In: *College teaching* 34.2 (1986), pp. 43–50 (cit. on p. 76).

[Fre87] Peter J Frederick. “Student involvement: Active learning in large classes”. In: *New Directions for Teaching and Learning* 1987.32 (1987), pp. 45–56 (cit. on p. 2).

[Gan+08] Gerald Gannod, Janet Burge, and Michael Helmick. “Using the inverted classroom to teach software engineering”. In: *2008 ACM/IEEE 30th International Conference on Software Engineering*. IEEE. 2008, pp. 777–786 (cit. on pp. 77–80).

[Gar+13] Martin Garbe, Jonas Vetterick, and Clemens H Cap. “Tweedback: online feedback system for large lectures”. In: *INFORMATIK 2013–Informatik angepasst an Mensch, Organisation und Umwelt* (2013), pp. 270–278 (cit. on p. 73).

[Geh01] Edward F Gehringer. “Electronic peer review and peer grading in computer-science courses”. In: *SIGCSE ’01: Proceedings of the thirty-second SIGCSE technical symposium on Computer Science Education*. 2001, pp. 139–143 (cit. on p. 120).

[Geh10] Edward F Gehringer. “Expertiza: Managing feedback in collaborative learning”. In: *Monitoring and Assessment in Online Collaborative Environments: Emergent Computational Technologies for E-Learning Support*. IGI Global, 2010, pp. 75–96 (cit. on pp. 119, 120).

[Gen35] Gerhard Gentzen. “Untersuchungen über das logische Schließen. I”. In: *Mathematische Zeitschrift* 39.1 (1935), pp. 176–210 (cit. on p. 38).

[Ges92] Joel Geske. “Overcoming the drawbacks of the large lecture class”. In: *College Teaching* 40.4 (1992), pp. 151–154 (cit. on p. 2).

[Gia+14] Michail N Giannakos, John Krogstie, and Nikos Chrisochoides. “Reviewing the flipped classroom research: reflections for computer science education”. In: *CSERC ’14: Proceedings of the Computer Science Education Research Conference*. 2014, pp. 23–29 (cit. on pp. 75, 79, 80).

[Gib92] Graham Gibbs. “Control and independence”. In: *Teaching Large Classes in Higher Education: How to Maintain Quality with Reduced Resources* (1992), pp. 37–59 (cit. on pp. 1, 2).

[Gie+12] Mark J Gierl, Hollis Lai, and Simon R Turner. “Using automatic item generation to create multiple-choice test items”. In: *Medical Education* 46.8 (2012), pp. 757–765 (cit. on p. 157).

[Gil+15] Mary Beth Gilboy, Scott Heinerichs, and Gina Pazzaglia. “Enhancing student engagement using the flipped classroom”. In: *Journal of Nutrition Education and Behavior* 47.1 (2015), pp. 109–114 (cit. on pp. 77–80, 82).

[Giu17] Luminiţa Giurgiu. “Microlearning an evolving elearning trend”. In: *Scientific Bulletin* 22.1 (2017), pp. 18–23 (cit. on p. 144).

[Gle86] Maryellen Gleason. “Better communication in large courses”. In: *College Teaching* 34.1 (1986), pp. 20–24 (cit. on pp. 1, 2).

[Glo+04] Ian Glover, Glenn Hardaker, and Zhijie Xu. “Collaborative annotation system environment (CASE) for online learning”. In: *Campus-Wide Information Systems* (2004), pp. 72–80 (cit. on pp. 15, 18).

[Got+10] Takuya Goto, Tomoko Kojiri, Toyohide Watanabe, Tomoharu Iwata, and Takeshi Yamada. “Automatic generation system of multiple-choice cloze questions and its evaluation”. In: *Knowledge Management & E-Learning: An International Journal* 2.3 (2010), pp. 210–224 (cit. on p. 157).

[GPI13] Edward F Gehringer and Barry W Peddycord III. “The inverted-lecture model: a case study in computer architecture”. In: *SIGCSE ’13: Proceeding of the 44th ACM technical symposium on Computer science education*. 2013, pp. 489–494 (cit. on pp. 77–80).

[Gra15] Colin Gray. “Designing online education for work based learners: Refining bite sized learning”. PhD thesis. Abertay University, 2015 (cit. on p. 143).

[Gra17] Krista Graham. “TechMatters: Peer to “Peergrade”: Exploring an Online Tool to Facilitate Peer Evaluation”. In: *LOEX Quarterly* 44.1 (2017), pp. 4–6 (cit. on p. 119).

[Gro17] Martin Gross. “Collective Peer Evaluation of Quiz Answers in Large Classes through Pairwise Matching”. Bachelor thesis. Institute of Informatics, LMU Munich, 2017 (cit. on p. 44).

[GT+13] Juan González-Tato, Martín Llamas-Nistal, Manuel Caeiro-Rodríguez, Fernando A Mikic-Fonte, et al. “Web-based audience response system using the educational platform called BeA”. In: *Journal of Research and Practice in Information Technology* 45.3/4 (2013), pp. 251–265 (cit. on pp. 29, 34).

[HA13] Matthias Hauswirth and Andrea Adamoli. “Teaching Java programming with the Informa clicker system”. In: *Science of Computer Programming* 78.5 (2013), pp. 499–520 (cit. on pp. 29, 30, 34, 35, 43).

[Hab05] Jacob Habgood. “Zombie Division: Intrinsic integration in digital learning games”. In: *Proceedings of the 8th Human Centred Technology Group Postgraduate Workshop* 576 (2005), pp. 45–48 (cit. on p. 179).

[Hal+10] Shivashankar Halan, Brent Rossen, Juan Cendan, and Benjamin Lok. “High score!-motivation strategies for user participation in virtual human development”. In: *International Conference on Intelligent Virtual Agents (IVA 2010)*. Springer. 2010, pp. 482–488 (cit. on p. 184).

[Hal96] Thomas M Haladyna. *Writing Test Items to Evaluate Higher Order Thinking*. Pearson, 1996 (cit. on p. 29).

[Ham+07] John Hamer, Catherine Kell, and Fiona Spence. “Peer assessment using Aropä”. In: *ACE ’07: Proceedings of the ninth Australasian conference on Computing education*. 2007, pp. 43–54 (cit. on pp. 119, 120).

[Ham+14] Juho Hamari, Jonna Koivisto, and Harri Sarsa. “Does gamification work?–a literature review of empirical studies on gamification”. In: *2014 47th Hawaii International Conference on System Sciences*. IEEE. 2014, pp. 3025–3034 (cit. on pp. 177, 183).

[Ham11] Juho Hamari. “Perspectives from behavioral economics to analyzing game design patterns: loss aversion in social games”. In: *CHI 2011 Social Games Workshop*. 2011 (cit. on pp. 198–201).

[Har18] Manuel Hartmann. “Reification for Backstage 2 – Refining Gamification to Give Learners a Tangible and Continuous Feedback”. Master thesis. Institute of Informatics, LMU Munich, 2018 (cit. on pp. 197, 201).

[Hat09] John Hattie. *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Routledge, 2009 (cit. on p. 2).

[Hau08] Matthias Hauswirth. “Informa: An extensible framework for group response systems”. In: *International Conference on Collaborative Computing: Networking, Applications and Worksharing*. Springer. 2008, pp. 271–286 (cit. on pp. 34, 35, 43).

[Hau11] Matthias Hauswirth. “Models and clickers for teaching computer science”. In: *7th Educators’ Symposium at MODELS*. 2011, pp. 41–44 (cit. on pp. 34, 35).

[Hel+18] Niels Heller, Sebastian Mader, and François Bry. “Backstage: A Versatile Platform Supporting Learning and Teaching Format Composition”. In: *Koli Calling ’18: Proceedings of the 18th Koli Calling International Conference on Computing Education Research*. ACM, 2018 (cit. on pp. xiv, 49).

[Hel+19] Niels Heller, Sebastian Mader, and François Bry. “More than the Sum of its Parts: Designing Learning Formats from Core Components”. In: *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*. ACM, 2019, pp. 2473–2476 (cit. on pp. xiv, 1, 49).

[Hel20] Niels Heller. “Pervasive Learning Analytics for Fostering Learner’s Self-Regulation”. PhD thesis. Institute for Informatics, LMU Munich, 2020 (cit. on pp. 1, 3, 4, 68, 206, 207, 214).

[Her+12] Michael J Herold, Thomas D Lynch, Rajiv Ramnath, and Jayashree Ramanathan. “Student and instructor experiences in the inverted classroom”. In: *2012 Frontiers in Education Conference Proceedings*. IEEE. 2012, pp. 1–6 (cit. on p. 78).

[HO14] David J Hornsby and Ruksana Osman. “Massification in higher education: Large classes and student learning”. In: *Higher Education* 67.6 (2014), pp. 711–719 (cit. on pp. 1, 2).

[Hof+09] Christian Hoff, Ulf Wehling, and Steffen Rothkugel. “From paper-and-pen annotations to artefact-based mobile learning”. In: *Journal of Computer Assisted Learning* 25.3 (2009), pp. 219–237 (cit. on pp. 15, 17, 28).

[How09] David C Howell. *Statistical methods for psychology*. Cengage Learning, 2009 (cit. on pp. 57, 89, 104, 149, 166, 190).

[HP94] Elizabeth Hazel and Michael Prosser. “First-year university students’ understanding of photosynthesis, their study strategies & learning context”. In: *The American Biology Teacher* 56.5 (1994), pp. 274–279 (cit. on p. 204).

[HS13] Clyde Freeman Herreid and Nancy A Schiller. “Case studies and the flipped classroom”. In: *Journal of College Science Teaching* 42.5 (2013), pp. 62–66 (cit. on p. 79).

[Hsi+16] Lu-Ho Hsia, Iwen Huang, and Gwo-Jen Hwang. “A web-based peer-assessment approach to improving junior high school students’ performance, self-efficacy and motivation in performing arts courses”. In: *British Journal of Educational Technology* 47.4 (2016), pp. 618–632 (cit. on p. 118).

[Hsu+08] Ying-Shao Hsu, Hsin-Kai Wu, and Fu-Kwun Hwang. “Fostering high school students’ conceptual understandings about seasons: The design of a technology-enhanced learning environment”. In: *Research in Science Education* 38.2 (2008), pp. 127–147 (cit. on p. 204).

[HT07] John Hattie and Helen Timperley. “The power of feedback”. In: *Review of Educational Research* 77.1 (2007), pp. 81–112 (cit. on p. 2).

[Hua+12] Yi-Ting Huang, Meng Chang Chen, and Yeali S Sun. “Personalized automatic quiz generation based on proficiency level estimation”. In: *20th International Conference on Computers in Education (ICCE 2012)*. 2012, pp. 553–560 (cit. on p. 156).

[Hul32] Clark L Hull. “The goal-gradient hypothesis and maze learning.” In: *Psychological Review* 39.1 (1932), pp. 25–43 (cit. on p. 199).

[Hun+16] Nathaniel J Hunsu, Olusola Adesope, and Dan James Bayly. “A meta-analysis of the effects of audience response systems (clicker-based technologies) on cognition and affect”. In: *Computers & Education* 94 (2016), pp. 102–119 (cit. on pp. 29, 31).

[Hun17] Aaron Chia Yuan Hung. “A critique and defense of gamification.” In: *Journal of Interactive Online Learning* 15.1 (2017), pp. 57–72 (cit. on p. 184).

[Hwa+07] Wu-Yuin Hwang, Chin-Yu Wang, and Mike Sharples. “A study of multimedia annotation of Web-based materials”. In: *Computers & Education* 48.4 (2007), pp. 680–699 (cit. on pp. 19, 20).

[Hwa+08] Wu-Yuin Hwang, Chin-Yu Wang, Gwo-Jen Hwang, Yueh-Min Huang, and Susan Huang. “A web-based programming learning environment to support cognitive development”. In: *Interacting with Computers* 20.6 (2008), pp. 524–534 (cit. on pp. 15, 120).

[Hwa+11] Wu-Yuin Hwang, Nian-Shing Chen, Rustam Shadiev, and Jin-Sing Li. “Effects of reviewing annotations and homework solutions on math learning achievement”. In: *British Journal of Educational Technology* 42.6 (2011), pp. 1016–1028 (cit. on pp. 15, 17, 19, 20, 27).

[Hwa14] Gwo-Jen Hwang. “Definition, framework and research issues of smart learning environments-a context-aware ubiquitous learning perspective”. In: *Smart Learning Environments* 1.1 (2014) (cit. on p. 41).

[Ima14] Jennifer Imazeki. “Bring-your-own-device: Turning cell phones into forces for good”. In: *The Journal of Economic Education* 45.3 (2014), pp. 240–250 (cit. on pp. 29, 34).

[Jen+14] Jamie L Jensen, Mark A McDaniel, Steven M Woodard, and Tyler A Kummer. “Teaching to the test... or testing to teach: Exams requiring higher order thinking skills encourage greater conceptual understanding”. In: *Educational Psychology Review* 26.2 (2014), pp. 307–329 (cit. on p. 30).

[JK09] Bethany C Johnson and Marc T Kiviniemi. “The effect of online chapter quizzes on exam performance in an undergraduate social psychology course”. In: *Teaching of Psychology* 36.1 (2009), pp. 33–37 (cit. on p. 145).

[Joh07] W Lewis Johnson. “Serious use of a serious game for language learning”. In: *Frontiers in Artificial Intelligence and Applications* 158 (2007), pp. 67–74 (cit. on p. 181).

[Jom+16] Omer Jomah, Amamer Khalil Masoud, Xavier Patrick Kishore, and Sagaya Aurelia. “Micro learning: A modernized education system”. In: *BRAIN. Broad Research in Artificial Intelligence and Neuroscience* 7.1 (2016), pp. 103–110 (cit. on p. 144).

[Jon+15] Anna Helga Jonsdottir, Audbjorg Jakobsdottir, and Gunnar Stefansson. “Development and Use of an Adaptive Learning Environment to Research Online Study Behaviour.” In: *Educational Technology & Society* 18.1 (2015), pp. 132–144 (cit. on p. 162).

[Kan16] Sean HK Kang. “Spaced repetition promotes efficient and effective learning: Policy implications for instruction”. In: *Policy Insights from the Behavioral and Brain Sciences* 3.1 (2016), pp. 12–19 (cit. on pp. 200, 201).

[Kar+06] Nikiforos Karamanis, Le An Ha, and Ruslan Mitkov. “Generating multiple-choice test items from medical text: A pilot study”. In: *INLG ’06: Proceedings of the Fourth International Natural Language Generation Conference*. 2006, pp. 111–113 (cit. on p. 157).

[Ke11] Fengfeng Ke. “A qualitative meta-analysis of computer games as learning tools”. In: *Gaming and Simulations: Concepts, Methodologies, Tools and Applications*. IGI Global, 2011, pp. 1619–1665 (cit. on pp. 177, 179, 181).

[KF05] Cem Kaner and Rebecca L Fiedler. “Inside out: A computer science course gets a makeover”. In: *Selected Papers On the Practice of Educational Communications and Technology Presented at The National Convention of the Association for Educational Communications and Technology* (2005), pp. 254–264 (cit. on pp. 77, 79, 80).

[Kib07] Jonathan Kibble. “Use of unsupervised online quizzes as formative assessment in a medical physiology course: effects of incentives on student participation and performance”. In: *Advances in Physiology Education* 31.3 (2007), pp. 253–260 (cit. on p. 145).

[Kim15] Bohyun Kim. “Designing Gamification in the Right Way”. In: *Library Technology Reports* 51.2 (2015), pp. 29–35 (cit. on pp. 177, 181, 183, 185).

[Kin93] Alison King. “From sage on the stage to guide on the side”. In: *College Teaching* 41.1 (1993), pp. 30–35 (cit. on pp. 75, 116, 117).

[Kiv+06] Ran Kivetz, Oleg Urminsky, and Yuhuang Zheng. “The goal-gradient hypothesis resurrected: Purchase acceleration, illusionary goal progress, and customer retention”. In: *Journal of Marketing Research* 43.1 (2006), pp. 39–58 (cit. on p. 199).

[KK01] José Kahan and Marja-Ritta Koivunen. “Annotea: an open RDF infrastructure for shared Web annotations”. In: *WWW ’01: Proceedings of the 10th international conference on World Wide Web*. 2001, pp. 623–632 (cit. on p. 17).

[KL09] Robin H Kay and Ann LeSage. “Examining the benefits and challenges of using audience response systems: A review of the literature”. In: *Computers & Education* 53.3 (2009), pp. 819–827 (cit. on pp. 29, 31).

[Koe+15] Kenneth R Koedinger, Jihee Kim, JuliZhuxin Jia, Elizabeth A McLaughlin, and Norman L Bier. “Learning is not a spectator sport: Doing is better than watching for learning from a MOOC”. In: *L@S ’15: Proceedings of the Second (2015) ACM Conference on Learning @ Scale*. ACM. 2015, pp. 111–120 (cit. on p. 145).

[Kov15] Geza Kovacs. “FeedLearn: Using facebook feeds for microlearning”. In: *Proceedings of the 33rd annual ACM Conference extended abstracts on human factors in computing systems*. ACM. 2015, pp. 1461–1466 (cit. on p. 144).

[Kra02] David R Krathwohl. “A revision of Bloom’s taxonomy: An overview”. In: *Theory into Practice* 41.4 (2002), pp. 212–218 (cit. on pp. 81, 82).

[Kös08] Sacit Köse. “Diagnosing student misconceptions: Using drawings as a research method”. In: *World Applied Sciences Journal* 3.2 (2008), pp. 283–293 (cit. on p. 204).

[Lag+00] Maureen J Lage, Glenn J Platt, and Michael Treglia. “Inverting the classroom: A gateway to creating an inclusive learning environment”. In: *The Journal of Economic Education* 31.1 (2000), pp. 30–43 (cit. on pp. 75, 77, 78).

[LE13] Kate Lockwood and Rachel Esselstein. “The inverted classroom and the CS curriculum”. In: *SIGCSE ’13: Proceeding of the 44th ACM technical symposium on Computer science education*. 2013, pp. 113–118 (cit. on pp. 77–79).

[Lec08] Thierry Lecomte. “Safe and reliable metro platform screen doors control/command systems”. In: *International Symposium on Formal Methods*. Springer. 2008, pp. 430–434 (cit. on p. 207).

[LH11] Joey J Lee and Jessica Hammer. “Gamification in education: What, how, why bother”. In: *Academic Exchange Quarterly* 15.2 (2011), pp. 1–5 (cit. on pp. 183, 185).

[Li+12] Wei Li, Tovi Grossman, and George Fitzmaurice. “GamiCAD: a gamified tutorial system for first time autocad users”. In: *UIST ’12: Proceedings of the 25th annual ACM symposium on User interface software and technology*. 2012, pp. 103–112 (cit. on pp. 185, 186).

[Lin06] Martin Lindner. “Use these tools, your mind will follow. Learning in immersive micromedia and microknowledge environments”. In: *The next generation: Research Proceedings of the 13th ALT-C Conference*. 2006, pp. 41–49 (cit. on p. 144).

[LK77] J Richard Landis and Gary G Koch. “The measurement of observer agreement for categorical data”. In: *Biometrics* 33.1 (1977), pp. 159–174 (cit. on p. 125).

[LL05] David G Lebow and Dale W Lick. “HyLighter: An effective interactive annotation innovation for distance education”. In: *20th Annual Conference on Distance Teaching and Learning*. 2005, pp. 1–5 (cit. on pp. 15, 17–19).

[LR09] Andrew Luxton-Reilly. “A systematic review of tools that support peer assessment”. In: *Computer Science Education* 19.4 (2009), pp. 209–232 (cit. on pp. 3, 118, 119).

[Mad+19] Sebastian Mader, Niels Heller, and François Bry. “Adding Narrative to Gamification and Educational Games with Generic Templates”. In: *Proceedings of the 18th European Conference on e-Learning (ECEL 2019)*. ACPI, 2019, pp. 360–368 (cit. on pp. xiv, 177, 199, 202, 203, 205, 206).

[Mad15] Sebastian Mader. “An Annotation Framework for a Collaborative Learning Platform”. Master thesis. Institute of Informatics, LMU Munich, 2015 (cit. on pp. 16–18, 27).

[Mah+13] Mary Lou Maher, Heather Lipford, and Vikash Singh. “Flipped classroom strategies using online videos”. In: *The Journal of Information Systems Education* 23.1 (2013), pp. 7–11 (cit. on p. 78).

[Mah+15] Mary Lou Maher, Celine Latulipe, Heather Lipford, and Audrey Rorrer. “Flipped classroom strategies for CS education”. In: *Proceedings of the 46th ACM Technical Symposium on Computer Science Education*. 2015, pp. 218–223 (cit. on pp. 77–80).

[Mai19] Anna Maier. “Adaptivity, Sequencing, and Scaffolding for a Web-based JavaScript Development Environment”. Master thesis. Institute of Informatics, LMU Munich, 2019 (cit. on pp. 36, 98, 101, 104).

[Mar07] Margie Martyn. “Clickers in the classroom: An active learning approach”. In: *EDUCAUSE Quarterly* 30.2 (2007), pp. 71–74 (cit. on pp. 2, 31, 32).

[Mar97] Catherine C Marshall. “Annotation: from paper books to the digital library”. In: *DL ’97: Proceedings of the second ACM international conference on Digital libraries*. 1997, pp. 131–140 (cit. on pp. 15, 20).

[MB13] Séamus McLoone and Conor Brennan. “A Smartphone-based Student Response System for Obtaining High Quality Real-time Feedback–Evaluated in an Engineering Mathematics Classroom: National University of Ireland Maynooth”. In: *Thinking Assessment in Science and Mathematics* (2013), pp. 148–154 (cit. on p. 29).

[MB18a] Sebastian Mader and François Bry. “Blending Classroom, Collaborative, and Individual Learning Using Backstage 2”. In: *8th International Conference in Methodologies and intelligent Systems for Technology Enhanced Learning (MIS4TEL 2018)*. Springer, 2018, pp. 3–11 (cit. on pp. xiii, 118, 139).

[MB18b] Sebastian Mader and François Bry. “Gaming the Lecture Hall: Using Social Gamification to Enhance Student Motivation and Participation”. In: *The Challenges of the Digital Transformation in Education - Proceedings of the 21st International Conference on Interactive Collaborative Learning (ICL2018)*. Springer, 2018, pp. 555–566 (cit. on pp. xiii, 177, 187).

[MB19a] Sebastian Mader and François Bry. “Audience Response Systems Reimagined”. In: *International Conference on Web-Based Learning (ICWL 2019)*. Springer, 2019, pp. 203–216 (cit. on pp. xiii, 29, 32–34, 42, 44).

[MB19b] Sebastian Mader and François Bry. “Fun and Engagement in Lecture Halls Through Social Gamification”. In: *International Journal of Engineering Pedagogy* 9.2 (2019), pp. 117–136 (cit. on pp. xiii, 177, 187–189, 191, 194, 195).

[MB19c] Sebastian Mader and François Bry. “Phased Classroom Instruction: A Case Study on Teaching Programming Languages”. In: *Proceedings of the 10th International Conference on Computer Supported Education*. SciTePress, 2019, pp. 241–251 (cit. on pp. xiii, 75, 88, 115).

[MB19d] Sebastian Mader and François Bry. “Towards an Annotation System for Collaborative Peer Review”. In: *International Conference in Methodologies and Intelligent Systems for Techhnology Enhanced Learning*. Springer, 2019, pp. 1–10 (cit. on pp. xiii, 117, 124, 125, 129, 130, 134, 140).

[MB20] Sebastian Mader and François Bry. “Promoting Active Participation in Large Programming Classes”. In: *Computers Supported Education*. Springer, 2020, to appear (cit. on pp. xiv, 75, 85, 88).

[McL+14] Jacqueline E McLaughlin, Mary T Roth, Dylan M Glatt, et al. “The flipped classroom: a course redesign to foster learning and engagement in a health professions school”. In: *Academic Medicine* 89.2 (2014), pp. 236–243 (cit. on pp. 77–80).

[Mey19] Maximilian Meyer. “A Browser-based Development Environmne for JavaScript Learning and Teaching”. Master thesis. Institute of Informatics, LMU Munich, 2019 (cit. on pp. 36, 83).

[Mil+60] George A Miller, Eugene Galanter, and Karl H Pribram. *Plans and the structure of behavior*. Henry Holt and Co, 1960 (cit. on p. 16).

[MK10] Catherine Mulryan-Kyne. “Teaching large classes at college and university level: Challenges and opportunities”. In: *Teaching in Higher Education* 15.2 (2010), pp. 175–185 (cit. on pp. 1, 2).

[MM17] Manuel Maarek and Léon McGregor. “Development of a Web Platform for Code Peer-Testing”. In: *The 8th Workshop on Evaluation and Usability of Programming Languages and Tools (PLATEAU) at SPLASH*. 2017 (cit. on p. 120).

[Mun11] Cristina Ioana Muntean. “Raising engagement in e-learning through gamification”. In: *Proceedings of the 6th International Conference on Virtual Learning*. 2011, pp. 323–329 (cit. on pp. 179, 186).

[Nah+14] Fiona Fui-Hoon Nah, Qing Zeng, Venkata Rajasekhar Telaprolu, Abhishek Padmanabhuni Ayyappa, and Brenda Eschenbrenner. “Gamification of education: a review of literature”. In: *International Conference on HCI in Business (HCIB 2014)*. Springer. 2014, pp. 401–409 (cit. on pp. 177, 182).

[Nat+04] Lasse Natvig, Steinar Line, and Asbjørn Djupdal. “‘Age of computers’: An innovative combination of history and computer game elements for teaching computer fundamentals”. In: *34th Annual Frontiers in Education*. IEEE. 2004 (cit. on p. 180).

[Nat+09] Lasse Natvig, Guttorm Sindre, and Asbjørn Djupdal. “A compulsory yet motivating question/answer game to teach computer fundamentals”. In: *Computer Applications in Engineering Education* 17.2 (2009), pp. 167–179 (cit. on pp. 180, 197).

[NC08] Joseph D Novak and Alberto J Cañas. *The theory underlying concept maps and how to construct and use them*. 2008 (cit. on pp. 203, 204).

[Nic+14] David Nicol, Avril Thomson, and Caroline Breslin. “Rethinking feedback practices in higher education: a peer review perspective”. In: *Assessment & Evaluation in Higher Education* 39.1 (2014), pp. 102–122 (cit. on p. 2).

[Nic10] David Nicol. “From monologue to dialogue: improving written feedback processes in mass higher education”. In: *Assessment & Evaluation in Higher Education* 35.5 (2010), pp. 501–517 (cit. on p. 2).

[Nic15] Scott Nicholson. “A recipe for meaningful gamification”. In: *Gamification in Education and Business*. Springer, 2015, pp. 1–20 (cit. on pp. 182, 183, 186, 197).

[Nik19] Stavros Nikou. “A micro-learning based model to enhance student teachers’ motivation and engagement in blended learning”. In: *Society for Information Technology & Teacher Education International Conference*. AACE. 2019, pp. 255–260 (cit. on p. 144).

[Nok+05] Petri Nokelainen, Miikka Miettinen, Jaakko Kurhila, Patrik Floréen, and Henry Tirri. “A shared document-based annotation tool to support learner-centred collaborative learning”. In: *British Journal of Educational Technology* 36.5 (2005), pp. 757–770 (cit. on p. 18).

[O’D+13] Siobhan O’Donovan, James Gain, and Patrick Marais. “A case study in the gamification of a university-level games development course”. In: *SAICSIT ’13: Proceedings of the South African Institute for Computer Scientists and Information Technologists Conference*. 2013, pp. 242–251 (cit. on pp. 185, 186, 197).

[Pet14] Dorian Peters. *Interface design for learning: Design strategies for learning experiences*. Pearson Education, 2014 (cit. on p. 23).

[PK13] Charles G Prober and Salman Khan. “Medical education reimagined: a call to action”. In: *Academic Medicine* 88.10 (2013), pp. 1407–1410 (cit. on pp. 1, 75).

[PL17] Carolyn M Plump and Julia LaRosa. “Using Kahoot! in the classroom to create engagement and active learning: A game-based technology solution for eLearning novices”. In: *Management Teaching Review* 2.2 (2017), pp. 151–158 (cit. on p. 187).

[Poh+12] Alexander Pohl, François Bry, Jeannette Schwarz, and Marlene Gottstein. “Sensing the classroom: Improving awareness and self-awareness of students in Backstage”. In: *2012 15th International Conference on Interactive Collaborative Learning (ICL)*. IEEE. 2012 (cit. on pp. 28, 72).

[Poh15] Alexander Pohl. “Fostering Awareness and Collaboration in Large-Class Lectures. Principles and Evaluation of the Backchannel Backstage”. PhD thesis. Institute for Informatics, Ludwig Maximilian University of Munich, 2015 (cit. on pp. 3, 4, 15, 20, 22, 23, 30, 51–56, 65, 66, 69, 72, 73, 196, 241–244).

[Pol04] George Polya. *How to solve it: A new aspect of mathematical method*. Princeton University Press, 2004 (cit. on p. 200).

[Pos+82] George J Posner, Kenneth A Strike, Peter W Hewson, and William A Gertzog. “Accommodation of a scientific conception: Toward a theory of conceptual change”. In: *Science Education* 66.2 (1982), pp. 211–227 (cit. on p. 204).

[Pou13] Malcom Poulin. “In learning, size matters”. In: *Chief Learning Officer* 12.2 (2013), pp. 38–56 (cit. on p. 144).

[PP15] Jan Papoušek and Radek Pelánek. “Impact of adaptive educational system behaviour on student motivation”. In: *International Conference on Artificial Intelligence in Education*. Springer. 2015, pp. 348–357 (cit. on p. 164).

[Pri04] Michael Prince. “Does active learning work? A review of the research”. In: *Journal of Engineering Education* 93.3 (2004), pp. 223–231 (cit. on p. 1).

[Qin+09] Jing Qin, Yim-Pan Chui, Wai-Man Pang, Kup-Sze Choi, and Pheng-Ann Heng. “Learning blood management in orthopedic surgery through gameplay”. In: *IEEE Computer Graphics and Applications* 30.2 (2009), pp. 45–57 (cit. on p. 181).

[Rad+93] Roy Rada, Sharon Acquah, Beverly Baker, and Phillip Ramsey. “Collaborative learning and the MUCH system”. In: *Computers & Education* 20.3 (1993), pp. 225–233 (cit. on p. 119).

[Rat+03] Matt Ratto, R Benjamin Shapiro, Tan Minh Truong, and William G Griswold. “The activeclass project: Experiments in encouraging classroom participation”. In: *Designing for Change in Networked Learning Environments*. Springer, 2003, pp. 477–486 (cit. on pp. 1–3).

[Ray11] Rick Raymer. “Gamification: using game mechanics to enhance eLearning”. In: *eLearn* 2011.9 (2011) (cit. on p. 198).

[Raz+12] Selen Razon, Jeannine Turner, Tristan E Johnson, Guler Arsal, and Gershon Tenenbaum. “Effects of a collaborative annotation method on students’ learning and learning-related motivation and affect”. In: *Computers in Human Behavior* 28.2 (2012), pp. 350–359 (cit. on pp. 15, 19).

[RB06] Robert J Roselli and Sean P Brophy. “Experiences with formative assessment in engineering classrooms”. In: *Journal of Engineering Education* 95.4 (2006), pp. 325–333 (cit. on p. 32).

[RB15] Sameen Reza and Maria Baig. “A study of inverted classroom pedagogy in computer science teaching”. In: *International Journal of Research Studies in Educational Technology* 4.2 (2015), pp. 19–30 (cit. on pp. 77–80).

[RC93] Peter J Rousseeuw and Christophe Croux. “Alternatives to the median absolute deviation”. In: *Journal of the American Statistical Association* 88.424 (1993), pp. 1273–1283 (cit. on pp. 57, 89, 104, 149, 166).

[Rea+18] Eliseo Reategui, Ana Paula M Costa, Daniel Epstein, and Michel Carniato. “Learning Scientific Concepts with Text Mining Support”. In: *International Conference in Methodologies and intelligent Systems for Techhnology Enhanced Learning (MIS4TEL 2018)*. Springer. 2018, pp. 97–105 (cit. on p. 204).

[Rei+96] Fraser JM Reid, Vlastimil Malinek, Clifford JT Stott, and Jonathan St BT Evans. “The messaging threshold in computer-mediated communication”. In: *Ergonomics* 39.8 (1996), pp. 1017–1037 (cit. on p. 22).

[RIK06] Henry L Roediger III and Jeffrey D Karpicke. “The power of testing memory: Basic research and implications for educational practice”. In: *Perspectives on Psychological Science* 1.3 (2006), pp. 181–210 (cit. on p. 145).

[Rob65] John Alan Robinson. “A machine-oriented logic based on the resolution principle”. In: *Journal of the ACM (JACM)* 12.1 (1965), pp. 23–41 (cit. on p. 36).

[Ros+18] Bella Ross, Anne-Marie Chase, Diane Robbie, Grainne Oates, and Yvette Absalom. “Adaptive quizzes to increase motivation, engagement and learning outcomes in a first year accounting unit”. In: *International Journal of Educational Technology in Higher Education* 15.1 (2018) (cit. on pp. 162, 163).

[RR10] Byron Reeves and J Leighton Read. “Ten ingredients of great games”. In: *Learning Circuits* (2010) (cit. on pp. 182, 186, 196).

[San+17] Robert Sanderson, Benjamin Young, and Paolo Ciccarese. *Web Annotation Data Model*. W3C Recommendation. https://www.w3.org/TR/2017/REC-annotation-model-20170223/. W3C, 2017 (cit. on p. 17).

[Sar12] Sukanta Sarkar. “The role of information and communication technology (ICT) in higher education for the 21st century”. In: *The Science Probe* 1.1 (2012), pp. 30–40 (cit. on pp. 2, 3).

[Sar14] Namita Sarawagi. “A flipped CS0 classroom: applying Bloom’s taxonomy to algorithmic thinking”. In: *Journal of Computing Sciences in Colleges* 29.6 (2014), pp. 21–28 (cit. on pp. 77–80, 82).

[SB11] Anna-Lise Smith and Lesli Baker. “Getting a clue: creating student detectives and dragon slayers in your library”. In: *Reference Services Review* (2011), pp. 628–642 (cit. on pp. 180, 197).

[Sch+11] Nancy M Schullery, Robert F Reck, and Stephen E Schullery. “Toward solving the high enrollment, low engagement dilemma: A case study in introductory business”. In: *International Journal of Business, Humanities and Technology* 1.2 (2011), pp. 1–9 (cit. on pp. 77, 78).

[Sch+15] Daniel Schön, Melanie Klinger, Stephan Kopf, Thilo Weigold, and Wolfgang Effelsberg. “Customizable learning scenarios for students’ mobile devices in large university lectures: a next generation audience response system”. In: *International Conference on Computer Supported Education (CSEDU 2015)*. Springer. 2015, pp. 189–207 (cit. on pp. 29, 34).

[Sch+16] Daniel Schön, Stephan Kopf, Melanie Klinger, and Benjamin Guthier. “New Scenarios for Audience Response Systems in University Lectures.” In: *13th International Conference on Cognition and Exploratory Learning in Digital Age (CELDA 2016)*. 2016, pp. 361–364 (cit. on p. 43).

[Sch07] Andreas Schmidt. “Microlearning and the Knowledge Maturing Process: Towards Conceptual Foundations for Work-Integrated Microlearning Support”. In: *Micromedia and Corporate Learning: Proceedings of the 3rd International Microlearning 2007 Conference*. IUP - Innsbruck University Press, 2007, pp. 99–105 (cit. on p. 144).

[Sch17] Maximilian Schwarzfischer. “Ein graphischer Editor zur Strukturierung multimedialer Dokumente”. Master thesis. Institute of Informatics, LMU Munich, 2017 (cit. on p. 10).

[Sch91] Martin D Schwartz. “Teaching the mass class: myths and tips”. In: *Journal of Criminal Justice Education* 2.2 (1991), pp. 255–266 (cit. on pp. 1, 2).

[SF15] Katie Seaborn and Deborah I Fels. “Gamification in theory and action: A survey”. In: *International Journal of Human-Computer Studies* 74 (2015), pp. 14–31 (cit. on pp. 177, 182).

[SH12] Kathrin F Stanger-Hall. “Multiple-choice exams: an obstacle for higher-level thinking in introductory science classes”. In: *CBE—Life Sciences Education* 11.3 (2012), pp. 294–306 (cit. on p. 29).

[She16] Kevin M Shea. “Beyond Clickers, Next Generation Classroom Response Systems for Organic Chemistry”. In: *Journal of Chemical Education* 93.5 (2016), pp. 971–974 (cit. on p. 29).

[Sin+08] Guttorm Sindre, Lasse Natvig, and Magnus Jahre. “Experimental validation of the learning effect for a pedagogical game on computer fundamentals”. In: *IEEE Transactions on Education* 52.1 (2008), pp. 10–18 (cit. on p. 180).

[SJ03] Jirarat Sitthiworachart and Mike Joy. “Web-based peer assessment in learning computer programming”. In: *Proceedings 3rd IEEE International Conference on Advanced Technologies*. IEEE. 2003, pp. 180–184 (cit. on p. 119).

[SM11] Liyan Song and Scot W McNary. “Understanding Students’ Online Interaction: Analysis of Discussion Board Postings.” In: *Journal of Interactive Online Learning* 10.1 (2011), pp. 1–14 (cit. on p. 184).

[SM12] Harald Søndergaard and Raoul A Mulder. “Collaborative learning through formative peer review: Pedagogy, programs and potential”. In: *Computer Science Education* 22.4 (2012), pp. 343–367 (cit. on p. 119).

[Sta+18] Marilyne Stains, Jordan Harshman, Megan K Barker, et al. “Anatomy of STEM teaching in North American universities”. In: *Science* 359.6383 (2018), pp. 1468–1470 (cit. on p. 2).

[Sta+19] Korbinian Staudacher, Sebastian Mader, and François Bry. “Automated Scaffolding and Feedback for Proof Construction: A Case Study”. In: *Proceedings of the 18th European Conference on e-Learning (ECEL 2019)*. ACPI, 2019, pp. 542–550 (cit. on pp. xiv, 36–38, 244).

[Sta18] Korbinian Staudacher. “Conception, Implementation and Evaluation of Proof Editors for Learning”. Bachelor thesis. Institute of Informatics, LMU Munich, 2018 (cit. on pp. 36, 38).

[Ste+10] Timothy Stelzer, David T Brookes, Gary Gladding, and José P Mestre. “Impact of multimedia learning modules on an introductory course on electricity and magnetism”. In: *American Journal of Physics* 78.7 (2010), pp. 755–759 (cit. on pp. 77–80).

[Ste16] Donald Stein. “Traumatic brain injury”. In: *Encyclopædia Britannica*. Encyclopædia Britannica, inc., 2016 (cit. on p. 146).

[Sto12] Bethany B Stone. “Flip your classroom to increase active learning and student engagement”. In: *Proceedings from 28th Annual Conference on Distance Teaching & Learning, Madison, Wisconsin, USA*. 2012 (cit. on pp. 77–80).

[Str12] Jeremy F Strayer. “How learning in an inverted classroom influences cooperation, innovation and task orientation”. In: *Learning Environments Research* 15.2 (2012), pp. 171–193 (cit. on pp. 77, 78).

[Su+10] Addison YS Su, Stephen JH Yang, Wu-Yuin Hwang, and Jia Zhang. “A Web 2.0-based collaborative annotation system for enhancing knowledge sharing in collaborative learning environments”. In: *Computers & Education* 55.2 (2010), pp. 752–766 (cit. on pp. 15, 19, 28).

[Tal13] Robert Talbert. “Learning MATLAB in the inverted classroom”. In: *The ASEE Computers in Education (CoED) Journal* 4.2 (2013), pp. 89–100 (cit. on pp. 77, 79).

[TG05] Des Traynor and J Paul Gibson. “Synthesis and analysis of automatic assessment methods in CS1: generating intelligent MCQs”. In: *ACM SIGCSE Bulletin* 37.1 (2005), pp. 495–499 (cit. on pp. 157, 158).

[TH04] John M Tauer and Judith M Harackiewicz. “The effects of cooperation and competition on intrinsic motivation and performance.” In: *Journal of Personality and Social Psychology* 86.6 (2004), pp. 849–861 (cit. on p. 187).

[Tha80] Richard Thaler. “Toward a positive theory of consumer choice”. In: *Journal of Economic Behavior & Organization* 1.1 (1980), pp. 39–60 (cit. on p. 199).

[Tin+13] David Tinapple, Loren Olson, and John Sadauskas. “CritViz: Web-based software supporting peer critique in large creative classrooms”. In: *Bulletin of the IEEE Technical Committee on Learning Technology* 15.1 (2013), pp. 29–35 (cit. on p. 118).

[Tip10] Christine D Tippett. “Refutation text in science education: A review of two decades of research”. In: *International Journal of Science and Mathematics Education* 8.6 (2010), pp. 951–970 (cit. on p. 204).

[TK91] Amos Tversky and Daniel Kahneman. “Loss aversion in riskless choice: A reference-dependent model”. In: *The Quarterly Journal of Economics* 106.4 (1991), pp. 1039–1061 (cit. on p. 198).

[Tod+17] Armando M Toda, Pedro HD Valle, and Seiji Isotani. “The dark side of gamification: An overview of negative effects of gamification in education”. In: *Researcher Links Workshop: Higher Education for All*. Springer. 2017, pp. 143–156 (cit. on pp. 177, 182–184).

[Top+00] Keith J Topping, Elaine F Smith, Ian Swanson, and Audrey Elliot. “Formative peer assessment of academic writing between postgraduate students”. In: *Assessment & Evaluation in Higher Education* 25.2 (2000), pp. 149–169 (cit. on p. 2).

[Top98] Keith Topping. “Peer assessment between students in colleges and universities”. In: *Review of Educational Research* 68.3 (1998), pp. 249–276 (cit. on pp. 2, 117, 138).

[Tor16] Carla Torgerson. “Bit by Bit”. In: *Talent Development* 70.11 (2016), pp. 26–28 (cit. on p. 144).

[Tra04] Stephan Trahasch. “From peer assessment towards collaborative learning”. In: *34th Annual Frontiers in Education*. IEEE. 2004 (cit. on p. 119).

[Tro99] Martin Trow. “From mass higher education to universal access: The American advantage”. In: *Minerva* (1999), pp. 303–328 (cit. on p. 1).

[Tur+16] Zeynep Turan, Zeynep Avinc, Kadir Kara, and Yuksel Goktas. “Gamification and education: Achievements, cognitive loads, and views of students”. In: *International Journal of Emerging Technologies in Learning (iJET)* 11.07 (2016), pp. 64–69 (cit. on pp. 185, 187).

[Val+86] Robert J Vallerand, Lise I Gauvin, and Wayne R Halliwell. “Negative effects of competition on children’s intrinsic motivation”. In: *The Journal of Social Psychology* 126.5 (1986), pp. 649–656 (cit. on p. 187).

[Var13] NV Varghese. “Governance reforms in higher education: A study of selected countries in Africa”. In: *Policy Forum on governance reforms in higher education in Africa*. UNESCO. 2013 (cit. on p. 1).

[Vil+14] Sergi Villagrasa, David Fonseca, and Jaume Durán. “Teaching case: applying gamification techniques and virtual reality for learning building engineering 3D arts”. In: *TEEM ’14: Proceedings of the Second International Conference on Technological Ecosystems for Enhancing Multiculturality*. 2014, pp. 171–177 (cit. on p. 186).

[VM+03] Jeroen JG Van Merriënboer, Paul A Kirschner, and Liesbeth Kester. “Taking the load off a learner’s mind: Instructional design for complex learning”. In: *Educational Psychologist* 38.1 (2003), pp. 5–13 (cit. on p. 1).

[Wan+16] Yanqing Wang, Yaowen Liang, Luning Liu, and Ying Liu. “A multi-peer assessment platform for programming language learning: considering group non-consensus and personal radicalness”. In: *Interactive Learning Environments* 24.8 (2016), pp. 2011–2031 (cit. on pp. 119–121).

[Wan15] Alf Inge Wang. “The wear out effect of a game-based student response system”. In: *Computers & Education* 82 (2015), pp. 217–227 (cit. on p. 187).

[Wan17] Simon Wanner. “Development and Evaluation of an Interface for Backstage 2 Focused on Good User Experience and Awareness”. Master thesis. Institute of Informatics, LMU Munich, 2017 (cit. on pp. 11, 12, 14, 20).

[Wes08] Peter S Westwood. *What teachers need to know about teaching methods*. ACER Press, 2008 (cit. on p. 3).

[Whi93] Edward M White. “Assessing higher-order thinking and communication skills in college graduates through writing”. In: *The Journal of General Education* 42.2 (1993), pp. 105–122 (cit. on p. 29).

[Wil13] Stephanie Gray Wilson. “The flipped class: A method to address the challenges of an undergraduate statistics course”. In: *Teaching of Psychology* 40.3 (2013), pp. 193–199 (cit. on pp. 77, 78, 80).

[WJ92] Andrew Ward and Alan Jenkins. “The problems of learning and teaching in large classes”. In: *Teaching Large Classes in Higher Education: How to Maintain Quality with Reduced Resources* (1992), pp. 23–36 (cit. on pp. 1, 2).

[WL16] Alf Inge Wang and Andreas Lieberoth. “The effect of points and audio on concentration, engagement, enjoyment, learning, motivation, and classroom dynamics using Kahoot”. In: *Proceedings of the 10th European Conference on Game Based Learning (ECGBL 2016)*. ACPI Limited. 2016, pp. 738–746 (cit. on pp. 186, 189).

[Woo+76] David Wood, Jerome S Bruner, and Gail Ross. “The role of tutoring in problem solving”. In: *The Journal of Child Psychology and Psychiatry* 17.2 (1976), pp. 89–100 (cit. on p. 1).

[Wou+13] Pieter Wouters, Christof Van Nimwegen, Herre Van Oostendorp, and Erik D Van Der Spek. “A meta-analysis of the cognitive and motivational effects of serious games.” In: *Journal of Educational Psychology* 105.2 (2013), pp. 249–265 (cit. on pp. 177, 179, 197).

[WR15] Lincoln C Wood and Torsten Reiners. “Storytelling to immersive learners in an authentic virtual training environment”. In: *Gamification in Education and Business*. Springer, 2015, pp. 315–329 (cit. on pp. 196, 197).

[Wri+15] James R Wright, Chris Thornton, and Kevin Leyton-Brown. “Mechanical TA: Partially automated high-stakes peer grading”. In: *Proceedings of the 46th ACM Technical Symposium on Computer Science Education*. 2015, pp. 96–101 (cit. on p. 117).

[Yar06] Sarita Yardi. “The role of the backchannel in collaborative learning environments”. In: *Proceedings of the 7th international conference on Learning sciences*. 2006, pp. 852–858 (cit. on p. 52).

[ZC11] Gabe Zichermann and Christopher Cunningham. *Gamification by design: Implementing game mechanics in web and mobile apps*. O’Reilly Media, Inc., 2011 (cit. on p. 183).

[ZW19] Jiahui Zhang and Richard E. West. “Designing Microlearning Instruction for Professional Development Through a Competency Based Approach”. In: *TechTrends* 64 (2019), 310—318 (cit. on p. 144).

[Bog11] Bogost, Ian. *Persuasive Games: Exploitationware*. `https://www.gamasutra.com/view/feature/134735/persuasive_games_exploitationware.php`, Last accessed on 2020-05-15. 2011 (cit. on p. 183).

[Eva09] Evan Miller. *How Not To Sort By Average Rating*. `https://www.evanmiller.org/how-not-to-sort-by-average-rating.html`, Last accessed on 2020-06-08. 2009 (cit. on p. 25).

[Joh04] John Gruber. *Markdown*. https://daringfireball.net/projects/markdown/, Last accessed on 2020-07-14. 2004 (cit. on p. 10).

[Kah20] Kahoot! *How to toggle points*. https://support.kahoot.com/hc/en-us/articles/115002303908, Last accessed on 2020-07-31. 2020 (cit. on p. 196).

[Mer20a] Merriam-Webster. *Fourth wall*. https://www.merriam-webster.com/dictionary/fourth%20wall, Last accessed on 2020-04-15. 2020 (cit. on p. 49).

[Mer20b] Merriam-Webster. *Reify*. https://www.merriam-webster.com/dictionary/fourth%20wall, Last accessed on 2020-04-24. 2020 (cit. on p. 198).

[Rob10] Robertson, Margaret. *Can't Play, Won't Play*. https://kotaku.com/cant-play-wont-play-5686393, Last accessed on 2020-05-15. 2010 (cit. on pp. 177, 182, 183).

[The16] The K!rew. *Boost collaboration with team mode in Kahoot!* https://kahoot.com/blog/2016/04/14/boost-collaboration-team-mode-kahoot/, Last accessed on 2020-05-27. 2016 (cit. on p. 187).

[Uni00] University College London. *Egyptian Chronology*. https://www.ucl.ac.uk/museums-static/digitalegypt//chronology/index.html, Last accessed on 2020-05-03. 2000 (cit. on pp. 158, 202).

Appendix

A

A.1 Large Class Lectures

A.1.1 Mapping Pohl's Constructs

In [Poh15], students' attitude towards Backstage was measured using the four constructs INTERACTIVITY, RATING, REWORK, and AWARENESS. To measure this, certain items from a total of 39 Likert items were selected to measure one of the aforementioned constructs. In the surveys conducted in the large classes described in this work, a similar survey was used to evaluate the use and effects of Backstage 2. The surveys used for that were similar, but not identical, and therefore, the four constructs could not be measured exactly in the way Pohl did. In the following, the Likert items selected by Pohl for each construct, and the equivalents used for the evaluations in this work are discussed.

INTERACTIVITY The construct INTERACTIVITY "measures the usefulness of Backstage as a means to promote interactivity in lectures" [Poh15, p. 68]. Table A.1 shows the statements Pohl used for measuring said construct and the equivalents used to measure the same construct in the surveys used in this work.

RATING The construct RATING "measures the students' assessments of rating to mark relevant backchannel comments" [Poh15, p. 68]. Table A.1 shows the statements Pohl used for measuring said construct and the equivalents used to measure the same construct in the surveys used in this work.

REWORK The construct REWORK "measures the usefulness of Backstage as a means to gather learning-related awareness" [Poh15, p. 68]. Table A.3 shows the statements Pohl used for measuring said construct and the equivalents used to measure the same construct in the surveys used in this work.

AWARENESS The construct AWARENESS " measures the usefulness of Backstage for reworking lectures" [Poh15, p. 68]. Table A.4 shows the statements Pohl used for

Tab. A.1.: Mapping of Pohl's Likert items measuring INTERACTIVITY to Likert items used in the surveys described in this work.

Statements from [Poh15]	Statements used here	Comment
Mir gefiel, in der Vorlesung öffentliche Kommentare mit Backstage erstellen zu können.	I liked to be able to create public comments on Backstage.	Direct translation
Es macht mir Spaß, die Quizfragen in Backstage zu beantworten.	–	No equvialence
Die Quizfragen sind ein geeignetes Mittel, Aktivität in der Vorlesung zu fördern.	Classroom quizzes are a suitable tool to make lectures more active.	Direct translation
Es hat mir Spaß gemacht, die Vorlesung zu besuchen und mit Backstage wurde die Vorlesung für mich angenehmer.	1. I had fun visiting the lecture. 2. Backstage made the lecture better than lectures without such a tool.	Split into two statements, second statement no direct translation
Mit Hilfe von Backstage habe ich viele Inhalte bereits in der Vorlesung verstanden.	Due to Backstage, I already understood much content during the lecture.	Direct translation
Die Quizfragen haben mir geholfen, zu erkennen, wo ich Probleme hatte.	–	No equivalence

Tab. A.2.: Mapping of Pohl's Likert items measuring RATING to Likert items used in the surveys described in this work.

Statements from [Poh15]	Statements used here	Comment
Mir gefiel, die Nachrichten auf Backstage bewerten zu können.	I liked to be able to up- and downvote comments on Backstage.	Direct translation
Die Bewertung der Nachrichten war geeignet um relevanten Nachrichten zu erkenne.	The displayed rating of messages was suitable for identifying relevant messages.	Direct translation
Es macht mir Spaß, Nachrichten in Backstage zu bewerten.	I had fun rating comments on Backstage.	Direct translation

Tab. A.3.: Mapping of Pohl's Likert items measuring REWORK to Likert items used in the surveys described in this work.

Statements from [Poh15]	Statements used here	Comment
Backstage ist für die Nachbereitung des Stoffs für die Vorlesungssitzungen nützlich gewesen.	1. The comments of my peers were useful for revising lectures. 2. The quizzes on Backstage were useful while revising the lecture content.	Split into two statements
Backstage ist für die Nachbereitung des Stoffs für die Übungen nützlich gewesen.	–	No equivalent
Backstage wird mir für die Nachbereitung des Stoffs für die Klausur hilfreich sein.	1. The comments created by my peers were useful while preparing for the examination. 2. The quizzes on Backstage were useful while preparing for the examination.	Split into two statements
Beim Wiederholen des Vorlesungsstoff war Backstage hilfreich.	–	No equivalent

measuring said construct and the equivalents used to measure the same construct in the surveys used in this work.

Tab. A.4.: Mapping of Pohl's Likert items measuring AWARENESS to Likert items used in the surveys described in this work.

Statements from [Poh15]	Statements used here	Comment
Ich finde es gut, dass ich auf der Übersichtsseite Informationen über mein Abschneiden in den Quizfragen bekomme.	I liked to get immediate feedback on my answers' correctness.	Similar, no direct translation.
Ich finde es gut, dass ich auf der Übersichtsseite Informationen über das Abschneiden meiner Kommilitonen bekomme.	I liked to be able to compare my results with the results of my peers.	Similar, no direct translation
Ich habe Backstage genutzt um zu erfahren, welche Fragen meine Kommilitonen haben.	I used Backstage to see my peer's questions.	Direct translation
Ich habe Backstage genutzt um die Antworten meiner Kommilitonen zu lesen.	I used Backstage to read my peer's answers.	Direct translation
Ich habe Backstage genutzt, um die Antworten der Tutoren und des Dozenten zu lesen.	I used Backstage to read the answers given by the teaching staff.	Direct translation

A.1.2 Survey

The following pages show the paper version of the survey used for the evaluation of **LC2** with the online version asking the same questions in a slightly different order. A nearly identical survey was used for the evaluation of **LC1** with two differences:

- The survey in **LC1** was conducted exclusively online.
- The online versions in both **LC1** and **LC2** additionally contained items referring to the problem-specific editors for the proof techniques Resolution and Natural Deduction. The results of these items in **LC1** were evaluated by Korbinian Staudacher and reported on in [Sta+19].
- The last four questions on the first page of the survey were added for **LC2** but referred to Backstage 2 / Projects, and hence, were evaluated by Niels Heller.

The System Usability Score was taken from the online version of the survey, which used a pre-defined version provided by SoSci Survey[1] which cite Brooke [Bro+96] as their source.

[1] https://www.soscisurvey.de/

Survey – „Logik und diskrete Strukturen“
Summer Term 2019

What is your course of study?

□ Informatics
□ Media Informatics
□ Bioinformatics
□ other _____

What semester are you in?

What is your gender?

□ Female □ Male □ Other

How old are you?

I skipped …

□ … no lectures. □ … at most two lectures. □ … more than two lectures.

I submitted homework in …

□ 10 to 13 of the weeks. □ 7 to 9 of the weeks. □ 4 to 6 of the weeks. □ 1 to 3 of the weeks. □ none of the weeks.

If you skipped homework. Why?

I received analytics reports per mail.

□ yes □ no

If you received analytics reports: The analytics reports ...

□ ... motivated me to learn more.
□ ... discouraged me.
□ ... motivated me to hand in the next assignments.
□ ... motivated me to learn more.
□ ... discouraged me.

Survey – „Logik und diskrete Strukturen“
Summer Term 2019

During lectures …

	yes	no	No answer
... I was logged into Backstage.	☐	☐	☐
... I created public comments on Backstage.	☐	☐	☐
... I created private comments on Backstage.	☐	☐	☐
... I commented on comments created by my peers on Backstage.	☐	☐	☐
... I up- and downvoted comments created by my peers on Backstage.	☐	☐	☐
... I participated in in-classroom quizzes.	☐	☐	☐

Outside of lectures …

	yes	no	No answer
... I visited Backstage.	☐	☐	☐
... I created public comments on Backstage.	☐	☐	☐
... I created private comments on Backstage.	☐	☐	☐
... I commented on comments created by my peers on Backstage.	☐	☐	☐
... I up- and downvoted comments created by my peers on Backstage.	☐	☐	☐
... I answered quizzes.	☐	☐	☐

Survey – „Logik und diskrete Strukturen“
Summer Term 2019

Functionalities of Backstage. *Please choose the appropriate response for each item.*

	strongly agree	agree	somewhat agree	somewhat disagree	disagree	strongly disagree
I liked to be able to create public comments on Backstage.	☐	☐	☐	☐	☐	☐
I liked to be able to create private comments on Backstage.	☐	☐	☐	☐	☐	☐
I liked to be able to up- and downvote comments on Backstage.	☐	☐	☐	☐	☐	☐
The displayed rating of messages was suitable for identifying relevant messages.	☐	☐	☐	☐	☐	☐
I had fun rating comments on Backstage.	☐	☐	☐	☐	☐	☐
I had fun visiting the lecture.	☐	☐	☐	☐	☐	☐
Backstage made the lecture better than lectures without such a tool.	☐	☐	☐	☐	☐	☐

Social Aspects of Backstage. *Please choose the appropriate response for each item.*

	strongly agree	agree	somewhat agree	somewhat disagree	disagree	strongly disagree
The lecture was suitable for the use of Backstage.	☐	☐	☐	☐	☐	☐
It was positive that the teaching staff participated in the exchange on Backstage	☐	☐	☐	☐	☐	☐
The lecturer sufficiently referred during lectures to the communication on Backstage.	☐	☐	☐	☐	☐	☐
I had the feeling that the communication on Backstage on Backstage gave the students more power of the flow of a lecture.	☐	☐	☐	☐	☐	☐
Without participation of the teaching staff in the exchange on Back-stage, Backstage would be much less useful.	☐	☐	☐	☐	☐	☐

Survey – „Logik und diskrete Strukturen“
Summer Term 2019

Your usage of Backstage. *Please choose the appropriate response for each item.*

	strongly agree	agree	somewhat agree	somewhat disagree	disagree	strongly disagree
I used Backstage to see my peer’s questions.	☐	☐	☐	☐	☐	☐
I used Backstage to read my peer’s answers.	☐	☐	☐	☐	☐	☐
I used Backstage to read the answers given by the teaching staff.	☐	☐	☐	☐	☐	☐
I liked to participate in the communication on Backstage.	☐	☐	☐	☐	☐	☐
The comments of my peers were useful for revising lectures.	☐	☐	☐	☐	☐	☐
The quizzes on Backstage were useful while revising the lecture content.	☐	☐	☐	☐	☐	☐
The comments created by my peers were useful while preparing for the examination.	☐	☐	☐	☐	☐	☐
The quizzes on Backstage were useful while preparing for the examination.	☐	☐	☐	☐	☐	☐
The dedicated editors for resolution and natural resolution were useful while preparing for the examination.	☐	☐	☐	☐	☐	☐
Due to Backstage, I already understood much content during the lecture.	☐	☐	☐	☐	☐	☐
Backstage distracted me from the lecture.	☐	☐	☐	☐	☐	☐
Notes on Backstage are no different from handwritten notes.	☐	☐	☐	☐	☐	☐
I had no or few incentives to participate in the communication on Backstage.	☐	☐	☐	☐	☐	☐

Survey – „Logik und diskrete Strukturen“
Summer Term 2019

The questions on this page refer *exclusively* to the quizzes conducted live during the during the lectures.

	strongly agree	agree	somewhat agree	somewhat disagree	disagree	strongly disagree
Classroom quizzes are a suitable tool to make lectures more active.	☐	☐	☐	☐	☐	☐
Instead of quizzes on Backstage, it would have been sufficient if the lecturer had asked questions orally and collected answers by counting raised hands.	☐	☐	☐	☐	☐	☐
Instead of quizzes on Backstage, it would have been sufficient if the lecturer had asked questions orally and afterwards provided the correct answers.	☐	☐	☐	☐	☐	☐
Quizzes helped me to re-focus my attention on lectures.	☐	☐	☐	☐	☐	☐
I liked to get immediate feedback on my answers’ correctness.	☐	☐	☐	☐	☐	☐
I liked to be able to compare my results with the results of my peers.	☐	☐	☐	☐	☐	☐
The overview of the classroom’s results helped me to better assess my knowledge.	☐	☐	☐	☐	☐	☐
The overview of the classroom’s results helped my to identify areas in which I was lacking understanding.	☐	☐	☐	☐	☐	☐
It would have been sufficient to only get feedback about the correctness of my own answer (without the classroom results).	☐	☐	☐	☐	☐	☐

Survey – „Logik und diskrete Strukturen“
Summer Term 2019

The questions on this page refer *exclusively* to the quizzes conducted live during the during the lectures.

	strongly agree	agree	somewhat agree	somewhat disagree	disagree	strongly disagree
Regardless of the team component, I would have participated in the quizzes.	☐	☐	☐	☐	☐	☐
I would have brought a device to participate in quizzes even without the team competition	☐	☐	☐	☐	☐	☐
I was motivated to participate in quizzes by the live overview of submitted quiz responses.	☐	☐	☐	☐	☐	☐
The competition with the other teams motivated me to participate in the quizzes.	☐	☐	☐	☐	☐	☐
I felt motivated to participate in quizzes, because my participation contributed to my team’s scores.	☐	☐	☐	☐	☐	☐

	strongly agree	agree	somewhat agree	somewhat disagree	disagree	strongly disagree
The teams made the lecture more engaging.	☐	☐	☐	☐	☐	☐
The competition with the other teams was fun.	☐	☐	☐	☐	☐	☐
I would have preferred to solve the quizzes on my own without points.	☐	☐	☐	☐	☐	☐
I would have preferred to solve the quizzes without a team but with points.	☐	☐	☐	☐	☐	☐
To improve my team’s chances of getting the quiz correct, I discussed the answer options with peers before giving an answer.	☐	☐	☐	☐	☐	☐

Survey – „Logik und diskrete Strukturen“
Summer Term 2019

Based on your experience with Backstage 2, do you agree with the statements below?
Please note that the *scale is reversed* in this question: strongly disagree is on the left side for this question.

	strongly disagree				strongly agree
I found the system unnecessarily complex.	☐	☐	☐	☐	☐
I found the system very cumbersome to use.	☐	☐	☐	☐	☐
I felt very confident using the system.	☐	☐	☐	☐	☐
I found the various functions in this system were well integrated.	☐	☐	☐	☐	☐
I thought the system was easy to use.	☐	☐	☐	☐	☐
I thought there was too much inconsistency in this system.	☐	☐	☐	☐	☐
I would imagine that most people would learn to use this system very quickly.	☐	☐	☐	☐	☐
I needed to learn a lot of things before I could get going with this system.	☐	☐	☐	☐	☐
I think that I would need the support of a technical person to be able to use this system.	☐	☐	☐	☐	☐
I think that I would like to use this system frequently.	☐	☐	☐	☐	☐

Survey – „Logik und diskrete Strukturen“
Summer Term 2019

	strongly agree	agree	somewhat agree	somewhat disagree	disagree	strongly disagree
Creating annotations was an intuitive process.	☐	☐	☐	☐	☐	☐
Annotations were presented in a clear manner.	☐	☐	☐	☐	☐	☐
While working with annotations, the filtering, ordering and sorting allowed me to maintain an overview on the annotations.	☐	☐	☐	☐	☐	☐
Being able to slide the list of annotations in and out is a good idea.	☐	☐	☐	☐	☐	☐
The quiz results were presented in a clear manner.	☐	☐	☐	☐	☐	☐
Giving an response to a quiz posed no problem.	☐	☐	☐	☐	☐	☐
The overview page of a single course had a clear design.	☐	☐	☐	☐	☐	☐
I had no problem to find the lecture material.	☐	☐	☐	☐	☐	☐
Backstage has an inviting layout.	☐	☐	☐	☐	☐	☐

Survey – „Logik und diskrete Strukturen“
Summer Term 2019

What are for you positive aspects of Backstage?

What are for you negative aspects of Backstage?

If you participated seldom or never on Backstage: What features would get you to use the platform more regularly?

Survey – „Logik und diskrete Strukturen“
Summer Term 2019

Do you think that Backstage is a tool to make large-class lectures more engaging? Why / Why not?

Do you think that quiz teams are suitable to make large-class lectures more engaging? Why / Why not?

A.2 Phased Classroom Instruction

A.2.1 Survey used in PCI1 and PCI2

Survey – Softwareentwicklungspraktikum „Spieleentwicklung mit JavaScript“
Winter Term 2018/19

What is your course of study?

- ☐ Informatics
- ☐ Media Informatics
- ☐ Bioinformatics
- ☐ other

What semester are you in?

What is your gender?

- ☐ Female
- ☐ Male
- ☐ Other

Which team are you in?

- ☐ Team A
- ☐ Team B
- ☐ Team C
- ☐ Team D

	strongly agree	agree	somewhat agree	somewhat disagree	disagree	strongly disagree
The immediate practical exercises after the mini lectures helped me understand the topic. Die sofortige praktische Übung nach den Mini-Vorlesungen hat mir beim Verstehen des Stoffs geholfen.	☐	☐	☐	☐	☐	☐
Discussions with my team mates during the practical exercises helped me understand the topic. Gespräche mit meinen Teammitgliedern während dem Lösen der Aufgaben haben mir beim Verstehen des Stoffes geholfen.	☐	☐	☐	☐	☐	☐
I would have preferred a traditional lecture without practical exercises. Ich hätte eine traditionelle Vorlesung ohne praktische Aufgaben bevorzugt.	☐	☐	☐	☐	☐	☐
I had fun during the plenum sessions. Ich hatte Spaß während den Plenumssitzungen.	☐	☐	☐	☐	☐	☐
Reviewing another team’s submission gave me *new* ideas where to improve my team’s submission. Das Bewerten von Lösungen anderer Teams hat mir *neue* Ideen gegeben, den Code des eigenen Teams zu verbessern.	☐	☐	☐	☐	☐	☐
The received review for our code helped me to identify weaknesses of my team’s code. Die Bewertung unserer Lösung durch andere Teams hat mir dabei geholfen, Schwächen im Code meines Teams zu identifizieren.	☐	☐	☐	☐	☐	☐

1

Survey – Softwareentwicklungspraktikum „Spieleentwicklung mit JavaScript“
Winter Term 2018/19

	strongly agree	agree	somewhat agree	somewhat disagree	disagree	strongly disagree
The mini lectures were sufficient to solve the practical exercises. Die Mini-Vorlesungen waren ausreichend um die Aufgaben zu lösen.	☐	☐	☐	☐	☐	☐
I would have preferred exercises that *do not* build upon each other. Ich hätte Aufgaben bevorzugt, die *nicht* aufeinander aufbauen.	☐	☐	☐	☐	☐	☐
The exercises were too difficult. Die Aufgaben waren zu schwer.	☐	☐	☐	☐	☐	☐
Through the mini lectures and practical exercises I feel well prepared for the implementation of the group project. Durch die Mini-Vorlesungen die Aufgaben fühle ich mich gut vorbereitet auf die Implementierung des Gruppenprojekts.	☐	☐	☐	☐	☐	☐
I liked that the exercises built upon each other. Ich fand es gut, dass die Aufgaben aufeinander aufgebaut haben.	☐	☐	☐	☐	☐	☐
The exercises were too big. Die Aufgaben waren zu umfangreich.	☐	☐	☐	☐	☐	☐

Survey – Softwareentwicklungspraktikum „Spieleentwicklung mit JavaScript“
Winter Term 2018/19

	strongly agree	agree	somewhat agree	somewhat disagree	disagree	strongly disagree
The JavaScript editor on Backstage made the getting started with JavaScript easy. Der JavaScript-Editor auf Backstage hat mir Einstieg in die JavaScript-Programmierung leicht gemacht.	☐	☐	☐	☐	☐	☐
The JavaScript-Editor was easy to operate. Der JavaScript-Editor war einfach zu bedienen.	☐	☐	☐	☐	☐	☐
The interface of Backstage, where exercises were worked on, was clearly designed. Die Ansicht auf Backstage, in der Aufgaben bearbeitet wurde, war übersichtlich gestaltet.	☐	☐	☐	☐	☐	☐
The interface of Backstage, where another team’s submission was reviewed, was clearly designed. Die Ansicht auf Backstage, in der die Abgabe eines anderen Teams bewertet wurde, war übersichtlich gestaltet.	☐	☐	☐	☐	☐	☐
The course format (i.e., mini lectures, followed by exercises and peer review) was well-supported by Backstage. Das Kursformat (d.h., Mini-Vorlesungen, gefolgt von Übungen und Peer Review) wurde gut von Backstage unterstützt.	☐	☐	☐	☐	☐	☐
I would have preferred to solve the practical exercises using a real development environment. Ich hätte es bevorzugt, die Aufgaben in einer echten Entwicklungsumgebung zu lösen.	☐	☐	☐	☐	☐	☐

Survey – Softwareentwicklungspraktikum „Spieleentwicklung mit JavaScript“
Winter Term 2018/19

What I liked most about the plenum ...
Was mir am meisten am Plenum gefallen hat ...

What could be done better in the future ...
Was man in Zukunft besser machen könnte ...

Other comments
Sonstige Kommentare

Survey – Softwareentwicklungspraktikum „Spieleentwicklung mit JavaScript“
Winter Term 2018/19

On a scale from 1 to 10, how do you estimate your programming experience at the start of the practical?

very inexperienced very experienced

☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐

How do you estimate your programming experience compared to your class mates at the start of the practical?

very inexperienced very experienced

☐ ☐ ☐ ☐ ☐

For how many years have you been programming?

How many courses did you take in which you had to write code (not including this practical)?

I wrote code outside of course’s assignments (“Übungsblätter”) in my free time.

☐ yes ☐ no

A.2.2 Survey used in PCI3 and PCI4

Survey – Softwareentwicklungspraktikum „Spieleentwicklung mit JavaScript“
Winter Term 2019/20

What is your course of study?

- ☐ Informatics
- ☐ Media Informatics
- ☐ Bioinformatics
- ☐ other

What semester are you in?

What is your gender?

- ☐ Female
- ☐ Male
- ☐ Other

	strongly agree	agree	somewhat agree	somewhat disagree	disagree	strongly disagree
The immediate practical exercises after the mini lectures helped me understand the topic. Die sofortige praktische Übung nach den Mini-Vorlesungen hat mir beim Verstehen des Stoffs geholfen.	☐	☐	☐	☐	☐	☐
Discussions with my team mates during the practical exercises helped me understand the topic. Gespräche mit meinen Teammitgliedern während dem Lösen der Aufgaben haben mir beim Verstehen des Stoffes geholfen.	☐	☐	☐	☐	☐	☐
I would have preferred a traditional lecture without practical exercises. Ich hätte eine traditionelle Vorlesung ohne praktische Aufgaben bevorzugt.	☐	☐	☐	☐	☐	☐
I had fun during the plenum sessions. Ich hatte Spaß während den Plenumssitzungen.	☐	☐	☐	☐	☐	☐
Reviewing another team’s submission gave me *new* ideas where to improve my team’s submission. Das Bewerten von Lösungen anderer Teams hat mir *neue* Ideen gegeben, den Code des eigenen Teams zu verbessern.	☐	☐	☐	☐	☐	☐
The received review for our code helped me to identify weaknesses of my team’s code. Die Bewertung unserer Lösung durch andere Teams hat mir dabei geholfen, Schwächen im Code meines Teams zu identifizieren.	☐	☐	☐	☐	☐	☐

1

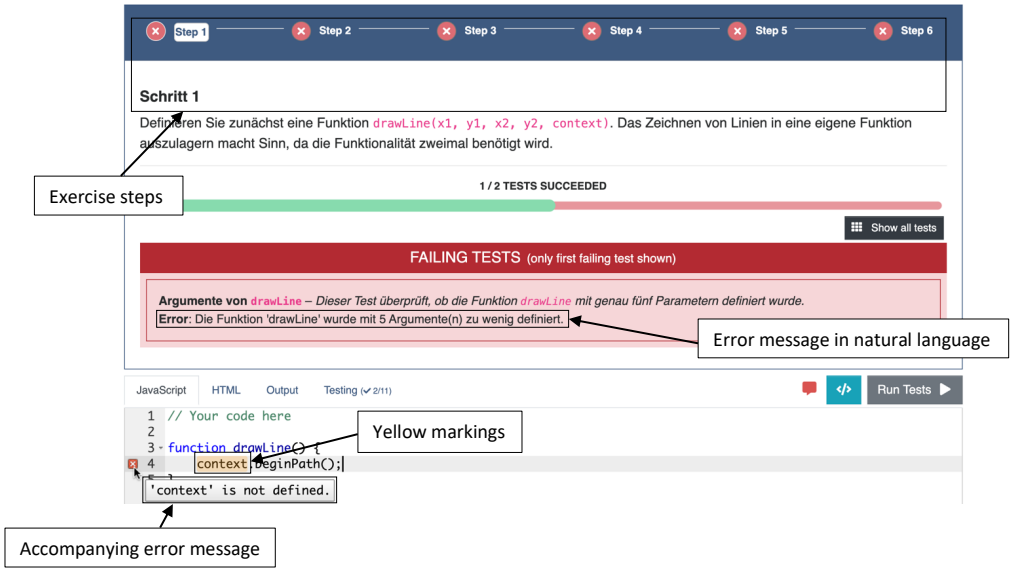
Survey – Softwareentwicklungspraktikum „Spieleentwicklung mit JavaScript“
Winter Term 2019/20

	strongly agree	agree	somewhat agree	somewhat disagree	disagree	strongly disagree
The mini lectures were sufficient to solve the practical exercises. Die Mini-Vorlesungen waren ausreichend um die Aufgaben zu lösen.	☐	☐	☐	☐	☐	☐
I would have preferred exercises that *do not* build upon each other. Ich hätte Aufgaben bevorzugt, die *nicht* aufeinander aufbauen.	☐	☐	☐	☐	☐	☐
The exercises were too difficult. Die Aufgaben waren zu schwer.	☐	☐	☐	☐	☐	☐
Through the mini lectures and practical exercises I feel well prepared for the implementation of the group project. Durch die Mini-Vorlesungen die Aufgaben fühle ich mich gut vorbereitet auf die Implementierung des Gruppenprojekts.	☐	☐	☐	☐	☐	☐
I liked that the exercises built upon each other. Ich fand es gut, dass die Aufgaben aufeinander aufgebaut haben.	☐	☐	☐	☐	☐	☐
The exercises were too big. Die Aufgaben waren zu umfangreich.	☐	☐	☐	☐	☐	☐
The lecturer was always there when my team had problems solving the exercise. Der Dozent war immer da, wenn mein Team beim Lösen der Aufgabe Probleme hatte.	☐	☐	☐	☐	☐	☐
The exercises were too easy. Die Aufgaben waren zu einfach.	☐	☐	☐	☐	☐	☐

Survey – Softwareentwicklungspraktikum „Spieleentwicklung mit JavaScript“
Winter Term 2019/20

	strongly agree	agree	somewhat agree	somewhat disagree	disagree	strongly disagree
The JavaScript editor on Backstage made the getting started with JavaScript easy. Der JavaScript-Editor auf Backstage hat mir Einstieg in die JavaScript-Programmierung leicht gemacht.	☐	☐	☐	☐	☐	☐
The JavaScript-Editor was easy to operate. Der JavaScript-Editor war einfach zu bedienen.	☐	☐	☐	☐	☐	☐
The interface of Backstage, where exercises were worked on, was clearly designed. Die Ansicht auf Backstage, in der Aufgaben bearbeitet wurde, war übersichtlich gestaltet.	☐	☐	☐	☐	☐	☐
The interface of Backstage, where another team’s submission was reviewed, was clearly designed. Die Ansicht auf Backstage, in der die Abgabe eines anderen Teams bewertet wurde, war übersichtlich gestaltet.	☐	☐	☐	☐	☐	☐
The course format (i.e., mini lectures, followed by exercises and peer review) was well-supported by Backstage. Das Kursformat (d.h., Mini-Vorlesungen, gefolgt von Übungen und Peer Review) wurde gut von Backstage unterstützt.	☐	☐	☐	☐	☐	☐
I would have preferred to solve the practical exercises using a real development environment. Ich hätte es bevorzugt, die Aufgaben in einer echten Entwicklungsumgebung zu lösen.	☐	☐	☐	☐	☐	☐

Survey – Softwareentwicklungspraktikum „Spieleentwicklung mit JavaScript“
Winter Term 2019/20



	strongly agree	agree	somewhat agree	somewhat disagree	disagree	strongly disagree
The yellow markings on the code and the accompanying error messages helped to identify error before running the code. Die gelben Markierungen auf dem Code und die dazugehörigen Fehlermeldungen halfen Fehler vor dem Ausführen des Codes zu finden.	☐	☐	☐	☐	☐	☐
It was clear to me at what point my team should move on to the next exercise step. Mir war klar, wann mein Team zum nächsten Schritt der Aufgabe wechseln sollte.	☐	☐	☐	☐	☐	☐
The error messages in natural language helped to recognize errors in my group’s code. Die Fehlermeldungen in natürlicher Sprache halfen Fehler im Code meiner Gruppe zu finden.	☐	☐	☐	☐	☐	☐
I would have preferred to see all failing tests instead of only one failing test. Ich hätte lieber alle scheiternden Tests eines Schritts anstatt nur einen scheiternden Test gesehen.	☐	☐	☐	☐	☐	☐

Survey – Softwareentwicklungspraktikum „Spieleentwicklung mit JavaScript“
Winter Term 2019/20

	strongly agree	agree	somewhat agree	somewhat disagree	disagree	strongly disagree
Switching between exercise steps was easy. Es war einfach zwischen den Schritten einer Aufgabe zu wechseln.	☐	☐	☐	☐	☐	☐
Even without the error messages in natural language, I would have been similar fast in recognizing errors in my group’s code. Auch ohne die Fehlermeldungen in natürlicher Sprache, wäre ich ähnlich schnell beim Finden der Fehler im Code meiner Gruppe gewesen.	☐	☐	☐	☐	☐	☐
Seeing only one exercise step helped focus solving that exercise step. Nur einen Schritt der Aufgabe zu sehen, half auf das Lösen dieses Schrittes zu konzentrieren.	☐	☐	☐	☐	☐	☐

Survey – Softwareentwicklungspraktikum „Spieleentwicklung mit JavaScript“
Winter Term 2019/20

What I liked most about the plenum ...
Was mir am meisten am Plenum gefallen hat ...

What could be done better in the future ...
Was man in Zukunft besser machen könnte ...

Other comments
Sonstige Kommentare

Survey – Softwareentwicklungspraktikum „Spieleentwicklung mit JavaScript“
Winter Term 2019/20

On a scale from 1 to 10, how do you estimate your programming experience at the start of the practical?

very inexperienced very experienced

☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐

How do you estimate your programming experience compared to your class mates at the start of the practical?

very inexperienced very experienced

☐ ☐ ☐ ☐ ☐

For how many years have you been programming?

How many courses did you take in which you had to write code (not including this practical)?

I wrote code outside of course’s assignments (“Übungsblätter”) in my free time.

☐ yes ☐ no

A.3 Collaborative Peer Review

A.3.1 Survey

The following section contains the paper version of the survey used for the evaluation of Collaborative Peer Review in **CPR5** and **CPR8**. In **CPR1-4**, an online survey was conducted which had minor differences to the paper version: Not in all venues the survey contained the System Usability Scale, and in the online versions, the questions measuring the students' attitudes towards the open access to essays and reviews were hidden behind a barrier question, that is, where only presented to the student when they previously confirmed to have looked at other students' essays or reviews. Another difference lies in the used scales: While the online surveys used a four-point Likert scale, the paper surveys used a six-point Likert scale. The scale was changed to raise more differentiated opinions.

The System Usability Score was taken from the online version of the survey, which used a pre-defined version provided by SoSci Survey[2] which cite Brooke [Bro+96] as their source.

[2] https://www.soscisurvey.de/

Survey – Masterseminar „Computational Ethics“
Summer Term 2019

What is your course of study?

□ Informatics □ Bioinformatics
□ Media Informatics □ other _____

What semester are you in?

What is your gender?

□ Female □ Male □ Other

How old are you?

1. During the seminar, the seminar papers were provided to all participants using Backstage. Each seminar paper was assigned two students for peer review. For peer review Backstage's annotation feature was used.

	yes	no
I looked at seminar papers other than those assigned to me for peer review while the peer review was running.	☐	☐
During the peer review, I up- and/or downvoted the other reviewer's annotations.	☐	☐
I examined only the reviews for my own seminar paper while the peer review was running.	☐	☐
During the peer review, I commented on the other reviewer's annotations.	☐	☐
I did **not** look at seminar papers other than my own while revising my own seminar paper.	☐	☐
I interacted (up/downvoting, commenting) with annotations on seminar papers other than the papers assigned to me for peer review.	☐	☐
I used the peer reviews created for my own seminar paper while revising my paper.	☐	☐

Survey – Masterseminar „Computational Ethics“
Summer Term 2019

2. Reviewing the seminar papers assigned to me ...

	strongly agree	agree	somewhat agree	somewhat disagree	disagree	strongly disagree
... gave me **new** ideas how to improve my own seminar paper.	☐	☐	☐	☐	☐	☐
... did **not** help me to get a better understanding of the standard of work in the course.	☐	☐	☐	☐	☐	☐
... was beneficial to my learning on aspects of scientific writing.	☐	☐	☐	☐	☐	☐
... allowed me – in contrast to other courses – to get a better assessment of my own performance.	☐	☐	☐	☐	☐	☐
... had few to none positive aspects.	☐	☐	☐	☐	☐	☐

3. The reviews I received from others ...

	strongly agree	agree	somewhat agree	somewhat disagree	disagree	strongly disagree
... helped me greatly to improve my seminar paper.	☐	☐	☐	☐	☐	☐
... were **not** beneficial for my learning on aspects of scientific writing.	☐	☐	☐	☐	☐	☐
... opened up new perspectives for me on how to write a seminar paper.	☐	☐	☐	☐	☐	☐
... had little to no use for me.	☐	☐	☐	☐	☐	☐
... offered more value than lecturer feedback.	☐	☐	☐	☐	☐	☐

Survey – Masterseminar „Computational Ethics“
Summer Term 2019

4. Answer the following block **only if** you looked at seminar papers besides those assigned to me for peer review.

	strongly agree	agree	somewhat agree	somewhat disagree	disagree	strongly disagree
Having access to all seminar papers helped me to assess my own performance.	☐	☐	☐	☐	☐	☐
I found aspects in other seminar papers (besides the papers assigned for peer review) which I used to improve my own seminar paper.	☐	☐	☐	☐	☐	☐
Having access to all seminar papers helped me to get a feeling for the standard of work in the course.	☐	☐	☐	☐	☐	☐
Having access to all seminar papers had little to no positive effects on my seminar paper.	☐	☐	☐	☐	☐	☐

5. Answer the following block **only if** you looked at peer reviews besides those created for my own seminar paper.

	strongly agree	agree	somewhat agree	somewhat disagree	disagree	strongly disagree
Looking at other participants' peer reviews gave me ideas for writing my own peer review.	☐	☐	☐	☐	☐	☐
I used suggestions made for other participants' seminar papers in improving my own paper.	☐	☐	☐	☐	☐	☐
Having access to all peer reviews had positive effects on the quality of my seminar paper.	☐	☐	☐	☐	☐	☐
Access to all peer reviews had no positive effects on my seminar paper.	☐	☐	☐	☐	☐	☐

Survey – Masterseminar „Computational Ethics“
Summer Term 2019

	yes	no
I did realize that I could choose between different content types for annotations.	☐	☐
I chose the appropriate content type for the majority of the annotations created by me.	☐	☐
I did realize that I could choose between highlighting a passage of text and creating a sticky note.	☐	☐

	strongly agree	agree	somewhat agree	somewhat disagree	disagree	strongly disagree
The predefined content types helped me to get a feeling for the areas that I should cover with the peer review.	☐	☐	☐	☐	☐	☐
Annotations were presented in a clear manner.	☐	☐	☐	☐	☐	☐
While viewing annotations, the filtering, ordering, and sorting means of Backstage made it easy to work with the annotations.	☐	☐	☐	☐	☐	☐
The number of up- and downvotes helped me to identify important annotations.	☐	☐	☐	☐	☐	☐
Creating an annotation was an intuitive process.	☐	☐	☐	☐	☐	☐

	strongly agree	agree	somewhat agree	somewhat disagree	disagree	strongly disagree
I think that peer review was a good fit for the course.	☐	☐	☐	☐	☐	☐
Giving peer review was too time-consuming.	☐	☐	☐	☐	☐	☐
I would have preferred a more traditional course design.	☐	☐	☐	☐	☐	☐
Receiving peer reviews from a single reviewer would have been sufficient.	☐	☐	☐	☐	☐	☐

6. Based on your experience with Backstage 2, do you agree with the statements below? Please note that the **scale has been reversed** for this question: strongly disagree is on the left side for this question!

	strongly disagree				strongly agree
I found the system unnecessarily complex.	☐	☐	☐	☐	☐
I found the system very cumbersome to use.	☐	☐	☐	☐	☐
I felt very confident using the system.	☐	☐	☐	☐	☐
I found the various functions in this system were well integrated.	☐	☐	☐	☐	☐
I thought the system was easy to use.	☐	☐	☐	☐	☐
I thought there was too much inconsistency in this system.	☐	☐	☐	☐	☐
I would imagine that most people would learn to use this system very quickly.	☐	☐	☐	☐	☐
I needed to learn a lot of things before I could get going with this system.	☐	☐	☐	☐	☐
I think that I would need the support of a technical person to be able to use this system.	☐	☐	☐	☐	☐
I think that I would like to use this system frequently.	☐	☐	☐	☐	☐

Survey – Masterseminar „Computational Ethics“
Summer Term 2019

What are for you notable positive aspects of the peer review using Backstage?

What are for you notable negative aspects of the peer review using Backstage?

Do you have suggestions or wishes for future versions of Backstage?

A.4 Bite-sized Learning

A.4.1 Survey

Welcome to the "Prüfungsvorbereitung Modul 4 (Neurologie)" - Survey

Thank you for using Backstage 2 for your preparation for the exam in Neurologie and thank you for taking part in this survey!

The following survey will help us gain an understanding of how to design courses such as the preparation course and how to improve Backstage 2. The survey will take about fifteen minutes.

Your answers are completely anonymous and cannot be traced back to the respondent.

Seite 02

Tell us something about yourself.

1. How old are you?

2. What is your gender?

- ◯ male
- ◯ female
- ◯ other

3. In what semester are you currently?

Seite 03

4. How much of the course did you complete?

How often did you retry an initally **correct** answered question?

How often did you retry an initially **incorrect** answered question?

If you never / rarely retried a question: Why?

- ◯ I had already seen the model solution.
- ◯ I did not know that I could redo answers.
- ◯ other, please specifiy

Seite 04

5. How often did you interrupt your work on the course?

- ◯ Never, did all sessions at once.
- ◯ Between sessions, did a single session at a time.
- ◯ Between sessions, did more than one session at a a time.
- ◯ other, please specify:

6. How did you work your way through a session?

- ◯ Question by question.
- ◯ Picked out only the questions I was interested in.
- ◯ Picked out only the question types I was interested in.
- ◯ other, please specify:

Seite 05

Below you find a number of statements about the usability of Backstage 2. Please read each statement carefully and decide to what extent you agree or disagree with the statement.

	strongly disagree	disagree	agree	strongly agree
Navigation from the start page of Backstage 2 to a session was intuitive.	○	○	○	○
Selecting a relevant area on an image was a straightforward process.	○	○	○	○
Starting a quiz was easy to do.	○	○	○	○
It was unclear how to operate Backstage 2.	○	○	○	○
It was not clear how to work through a session.	○	○	○	○

Seite 06

How helpful were the different question types for your exam preparation?

	not helpful at all	rather not helpful	helpful	extremly helpful
Open answer questions (“Freitextfrage”)	○	○	○	○
Scale questions	○	○	○	○
Marking relevant selections	○	○	○	○
Multiple choice questions	○	○	○	○

How clear was the process of submitting an answer for the different question types?

	unclear	quite unclear	quite clear	clear
Open answer question (“Freitextfrage”)	○	○	○	○
Scale question	○	○	○	○
Marking relevant sections	○	○	○	○
Multiple choice question	○	○	○	○

Seite 07

The explanation text which was provided after submitting an answer was ...

unhelpful	not very helpful	helpful	extremely helpful
○	○	○	○

How understandable was the feedback about the correct and incorrect answers for the different question types?

	unclear	quite unclear	quite clear	clear
Open answer question (“Freitextfrage”)	○	○	○	○
Scale question	○	○	○	○
Marking relevant selections	○	○	○	○
Multiple choice question	○	○	○	○

Seite 08

The comparison with my fellow students was helpful to assess my current learning progress.

strongly disagree	disagree	agree	strongly agree
○	○	○	○

I would have liked to get more information about the progress of my fellow students.

strongly disagree	disagree	agree	strongly agree
○	○	○	○

Seite 09

Below you find a number of statements about the design of the exam preparation course. Please read each statement carefully and decide to what extent you agree or disagree with the statement.

	strongly disagree	disagree	agree	strongly agree
The number of questions per session should have been bigger.	○	○	○	○
The mix of different question types within a session provided good variety.	○	○	○	○
The number of questions per session should have been smaller.	○	○	○	○
The overall number of questions was well-chosen.	○	○	○	○
Arranging all questions in a single session would have been better.	○	○	○	○
The number of questions per session was appropriate.	○	○	○	○

7. Below you find a number of statements about the use of Backstage 2 in the exam preparation course. Please read each statement carefully and decide to what extent you agree or disagree with the statement.

	strongly disagree	disagree	agree	strongly agree
The exam images provided on Backstage 2 were helpful for my exam preparation.	○	○	○	○
The content (besides the exam images) on Backstage 2 was helpful for my exam preparation.	○	○	○	○
Backstage 2 offered additional value (besides the exam images) not provided by any other e-learning software.	○	○	○	○
Using Backstage 2 for an exam preparation course was a good idea.	○	○	○	○
Any other e-learning software would have offered the same value as Backstage 2.	○	○	○	○
The variety of question types provided by Backstage 2 is not provided by any other e-learning software.	○	○	○	○

I would have used Backstage 2 and worked through a similar amount of content even if no exam images had been involved.

○ unlikely ○ rather unlikely ○ likely ○ very likely

Seite 11

What did you like most about Backstage 2?

What did you not like about Backstage 2?

What could be done better in the future?

Letzte Seite

Vielen Dank für Ihre Teilnahme!

Wir möchten uns ganz herzlich für Ihre Mithilfe bedanken.

Ihre Antworten wurden gespeichert, Sie können das Browser-Fenster nun schließen.

A.5 Social Gamification based on Teams

A.5.1 Survey

Survey on the use of Backstage 2 in Softwareentwicklungspraktikum „Spieleentwicklung mit JavaScript“

What is your course of study?

- ☐ Informatics
- ☐ Media Informatics
- ☐ Bioinformatics
- ☐ other

What semester are you in?

What is your gender?

- ☐ Female
- ☐ Male
- ☐ Other

	strongly agree	agree	disagree	strongly disagree
Regardless of the team component, I would have participated in quizzes.	☐	☐	☐	☐
I would have brought a device to participate in quizzes even without the team competition.	☐	☐	☐	☐
I was motivated to participate in quizzes by the live overview of submitted quiz responses.	☐	☐	☐	☐
The competition with other teams motivated me to give answers to quizzes.	☐	☐	☐	☐
I felt motivated to participate in quizzes, because my participation contributed to the team score.	☐	☐	☐	☐

1

Survey on the use of Backstage 2 in Softwareentwicklungspraktikum „Spieleentwicklung mit JavaScript“

	strongly agree	agree	disagree	strongly disagree
The use of teams made the lecture more engaging.	☐	☐	☐	☐
To improve my team’s chances of getting the quiz correct, I discussed the answer options to quizzes with my team before giving an answer.	☐	☐	☐	☐
The competition with the other teams was fun.	☐	☐	☐	☐
I tried the quiz’ code (if possible) to improve my team’s chances of getting the quiz correct.	☐	☐	☐	☐
I would have preferred to solve the quizzes on my own without points.	☐	☐	☐	☐
I would have preferred to solve the quizzes without a team but with points.	☐	☐	☐	☐

What I most liked about the team component and the quizzes …

What could be done better in the future?

Colophon

This thesis was typeset with LaTeX 2$_\varepsilon$. It uses the *Clean Thesis* style developed by Ricardo Langner. The design of the *Clean Thesis* style is inspired by user guide documents from Apple Inc.

Download the *Clean Thesis* style at `http://cleanthesis.der-ric.de/`.