ABSTRACT
Teaching a mandatory course for undergraduate computer science students with up to 750 students per semester, we have been making extensive use of peer reviewing. During the semester, each student has to work on a set of assignments. After finishing an assignment, the student has to write three peer reviews for three pieces of work by different, anonymous peers. One of the problems in the use of (student) peer reviewing in large university courses is the quality of written reviews. To address this problem, we devised various provisions to maintain or increase reviewing quality. In this article, we describe one of these provisions, namely the use of three different types of reviews instead of using identical review types three times in a row: guided reviews, open reviews, checkbox reviews. Our aim in this article is to research the impact of these different review types on the students’ experience and acceptance of the reviewing process to inform the design of a reviewing process that better fits the students’ needs. To gain such insights, we gathered feedback using a survey, which was completed by 101 students. Using qualitative analysis, we extracted and defined room for improvement and discuss possible changes for our current peer review system and process. Our learnings show insights into the types of reviews students prefer, and hint at some advantages and pitfalls of peer reviewing that can have substantial impact on the design and application of such a system in large university courses.

CCS CONCEPTS
• Human-centered computing → Human computer interaction (HCI).

KEYWORDS
peer review, peer feedback, technology-enhanced learning, computer science education

1 INTRODUCTION
In a course for first semester computer science students we make extensive use of anonymous peer reviewing. Even though the course concept itself has undergone a recent overhaul [13], we have been using this organisational form for more than ten years now. Over this period, we have continuously refined the platform we use to organise peer reviewing [26], based on feedback and with the participation of our students [28].

The handling of exercises - issuing assignments and accepting hand-ins - is organised online. This enables us to have students review each others’ work on a massive scale. With up to 750 participants, we facilitate up to 40,000 individual peer reviews each semester. The allocation and organisation of these reviews happens automatically, using a bespoke software platform we developed over the years.

Applying our usual process of evaluation and redesign based on data collected at the end of a semester, we regularly introduce new features, many of which are targeted to increase the quality of student peer reviewing [27]. Inspired and informed by feedback from students, we recently changed our system so that reviewing becomes a more varied work: Instead of reviewing the work of three peers following identical review questions, we introduced review types, so that the three reviews students have to write each concern different aspects and have separate sets of review questions.

This paper discusses the results of a survey we conducted in the winter semester of 2020 to evaluate the success of our redesign, and to inform the further development of the course as well as the software platform.

2 STATE OF THE ART
Large courses create a demanding context for reviewing. The number of participants in such a course renders expert reviewing unsustainable (e.g. [38, 40]). Peer reviewing provides an instrument to avoid this problem. Consequently, a sizeable body of research has been produced on peer feedback and reviewing (e.g. [18, 22, 25]). The use of peer reviewing has been described in various learning contexts such as game design courses (e.g. [12, 40]), computer science courses (e.g. [1, 24]), psychology and writing courses (e.g. [4, 10]), lab reports (e.g. [5, 29]), and coding and design exercises (e.g. [41]).

In the context of this work, it is important to differentiate between peer review, feedback and assessment (further also referred to as evaluation). Review describes the process of reading a submission and writing qualitative feedback, potentially complemented with a quantitative assessment. Feedback is the actual written text provided to help authors to improve their work. Studies show that
feedback is effective by being specific, actionable and justified [33], as well as timely, motivational, individual and manageable [16]. Assessment is the evaluation of said work on a given scale.

The importance of a clean distinction between feedback and assessment has been shown by Hicks et al.: “A reviewer who perceives their task to be assigning a final grade may give critical feedback that focuses on identifying flaws. By contrast, a reviewer who perceives their task to be helping a learner improve may provide more explanations and suggest changes to the reviewer.” [18, p.459]. Since in the context of this paper, we discuss peer review and peer feedback while assessment and evaluation is done by the course team, this differentiation is valuable in defining our goals and impacts the perceived quality of a review.

While peer reviewing seems to offer improvements in the organisation and execution of large courses, it has definitely shown positive effects on the learning process itself (e.g. [19, 32]). Further studies have shown that peer reviewing builds problem solving skills and encourages reflection, self awareness [35] and critical thinking [21]. Furthermore, it has been shown to improve the performance in exams [36], to widen students’ horizons by confronting them with different points of views and strategies for the solution of similar assignments [12, 36], and to eliminate misconceptions early in the learning process [14]. By exposing students to different perspectives, they learn to grasp the intricacies of a topic, deepen the subject knowledge [12, 42], and generate new insights [32].

Feedback is an informal assessment tool that provides timely, descriptive information regarding direct observations of the learner in the learning environment [20]. It has been described as a key mechanism in the development and enhancement of learning [31]. Brown 2001 noted that feedback is best if it is timely, perceived as relevant, meaningful, and suggests ways of improvement which are within the student’s grasp [8]. Jug et al. realise that “we are all simultaneously educators and learners, both giving and receiving feedback on a daily basis. Thus, competence in giving and receiving feedback is crucial […]” [20].

Feedback sequencing and timing came under closer scientific scrutiny only recently. Henley and Reed [17] note that the feedback sandwich method of layering the corrective feedback between two layers of appreciative feedback is often recommended, but it’s rarely proven. Several studies have show the feedback sandwich approach to be less than optimal [34, 43], while others have shown evidence of its effectiveness [37]. On the other hand, a consensus of the positive impact of writing feedback can be found. One of the more interesting observations was that “equipping learners to engage in feedback processes may reduce the emotional burden on both parties, rendering techniques such as the feedback sandwich redundant.” [30]. They suggest viewing feedback as a relational activity.

Different types of feedback in peer reviewing have been subject to research before. Studies have been conducted on numeric feedback (e.g. [11, 18]), rubric feedback (e.g. [2, 22]), textual feedback (e.g. [15, 21]), and visual feedback (e.g. [1, 6]), among others. Numeric and rubric feedback were described as being less time-intensive, easier to compare across reviews and also easier to understand [11], at the cost of lacking depth and being less helpful in the improvement of the reviewed work [18].

Textual feedback takes more time to produce [23], which can impact student motivation and distract from other tasks in the course, or even outside of the course [3]. Yet, textual feedback was found to be more useful [22], especially when based on structured questions (e.g. [15, 18]), when students had a chance to pose their own questions [4], or when triggers and prompts were used to elucidate constructive feedback [21].

Visual feedback has been used to directly point to problematic parts of submissions with colour coding to label them as correct, wrong or irrelevant [1], giving authors a quick overview of the parts of their work that might need revision. Combining review types have also been studied, e.g. textual along with numeric feedback (e.g. [18]). Some studies focused on comparative reviewing, comparing either different drafts of the same submission, thus showcasing the author’s progress [18], or between works from different authors in order [9].

3 PEER REVIEWING

Our bespoke reviewing software, Aurora [26], is an online learning management system designed and implemented [28] over many years to meet our requirements and needs. We use it to organise and conduct a first-semester course Ways of Thinking in Informatics [13] that is mandatory for all computer science undergraduate programs at TU Wien. For their grade, students hand in exercises on six of the main chapters of the course. In a nod toward game-based learning, we call these exercises challenges. Each challenge is made up of typically two to four tasks that have to be completed sequentially. After each task, participants are asked to review three elaborations from other students for the just completed task. Having completed these reviews, students can then work on the next task in the challenge. The final task in every challenge is called Feedback & Reflection and is not subject to reviewing. We use the final task to have students reflect on their learnings, and gather feedback for the improvement of the challenges for consecutive semesters.

The process is best explained with an example; in the first task of a challenge on algorithmic bias, a student would have to research on the origins of algorithmic bias, based on a series of guiding questions, and write up a short report as a result. When the author hands in their work, it is added to the review pool, so that other students can act as reviewers for this elaboration. At the same time, three already handed-in task elaborations by other students are assigned to the author to write three different reviews. After reviewing these elaborations, the author can go on to the second task, where they analyse a training data set of portrait pictures regarding data bias. Again, the author hands in their work to be reviewed and has to write three reviews for work handed in by other students.

Similar to the assignment in this example, the challenges we pose from various problem domains that are typically associated with the term wicked problems [39]. Our exercises usually have no right or wrong answers, but dabble in fields where reflective practice, differentiated views and deliberative discussion are asked for, rather than discovering or finding the right answer.

We guide students by offering them how-to descriptions of scientific writing, good reviewing, avoiding and checking for plagiarism, accepting and passing on critique, with a special emphasis on the
feedback sandwich as a simple formula to write reviews that can be accepted more readily.

After handing in the final task of a challenge, the accumulated task elaborations are evaluated by the course team, which usually includes course organisers, lecturers and tutors selected from students who have previously excelled in the course). Each student receives a statement as formative feedback for their work, a number of points as summative evaluation, with possible bonus points for extraordinary quality, and a rating on a quadrinomial scale for the quality of each of their written reviews. A cumulative score derived from this quality rating is used as a make-or-break criterion at the end of the semester: students whose score is below a threshold value will fail the course. Furthermore, students can rate reviews on the same scale, and see the review ratings they received from the course team as well as from their peers.

3.1 Reviewing types

As can be seen in Fig.1, all types of reviews have some elements in common; they include a text entry addressed to the course team rather than the author of the reviewed work in case the reviewer wants to share some thoughts privately. Each review concludes with a Final Feedback section where reviewers have to assess the overall quality of the reviewed work with one of four carefully worded options.

We follow a couple of guidelines in writing review questions for this type of reviews; questions should be phrased as if the author asks the reviewer a question; they should be phrased so that they cannot be answered with a simple yes or no; start with more general questions, followed by more specific ones; asking for critique should include asking for suggested improvements. Since reviewers have just finished the same task before they started reviewing, questions can also ask to compare the work of the author to the work of the reviewer, to point out differences or similarities.

3.1.2 Open Reviews. In open reviews, students sum up whatever they think is necessary as feedback for the author so they can improve their work. These reviews are always led by a simple pair of questions – “What do you like about my work?” and “What could I do to improve my work?”. Along with these two questions, we suggest students apply the strategy of writing feedback in form of a feedback sandwich, layering positive and negative remarks to make the feedback more acceptable.

3.1.3 Checkbox Reviews. For checkbox reviews, the requirements of a task are represented by checkboxes or radio buttons. First, reviewers have to check a list of individual requirements derived from the assignment text. Each item is displayed as a checkbox, and represents an expected requirement for the work that is reviewed. Reviewers have to examine whether this requirement is covered by the reviewed work, thereby checking the reviewed work for completeness.

Additionally, checkboxes concerning formal requirements such as correct use of references, use of adequate academic language, formatting and layout are included, and a basic plagiarism check should be conducted. These questions are answered using pairs of radio buttons (see Fig.1 for examples) offering the options rather yes and rather no to avoid overly restrictive interpretation. These requirements are the same for most assignments. When individual requirements do not make sense – such as correct use of references when no references are expected – the corresponding question is dropped. Finally, checkbox reviews include a text entry for optional miscellaneous remarks addressed to the author of the reviewed work.

A list of review questions from all review types is available to the author while they write their original elaboration as a section of the assignment titled “Your peers will review your work by answering the following questions”.

3.1.1 Guided Reviews. Guided reviews were the original type of reviews. We modelled this review type after conference peer review systems where reviewing is scaffolded by a series of questions targeting different aspects of quality.

Guided reviews consist of two or more predefined questions that can be answered in medium-sized text fields. The questions differ adapting to the content of the respective task. Most of these questions are mandatory, but optional questions are possible. Additionally, guided reviews include a text entry for optional miscellaneous remarks addressed to the author of the reviewed work.

4 SURVEY RESULTS

The survey was conducted towards the end of the winter semester 2020. 101 of 670 students participated in the voluntary survey after having used the reviewing system for the previous 4 months in a distance learning setting. The survey was conducted using a Lime Survey instance, which was hosted locally on servers run by our research group. All data was anonymised before the evaluation.

4.1 Structure and Analysis

The survey consisted of fourteen mainly open questions about the review system. Only seven of these questions focused primarily on types of reviews, reviewing issues and how reviewing could
be enhanced in future iterations of the system. Subsequently, only these seven questions were evaluated for this article.

The evaluation of the survey was done using the qualitative approach of thematic analysis [7]. We evaluated each question by first summarising the content in initial codes, iteratively reworking these codes until they fit the types of answers available for each question. When all answers were coded, we looked for overarching themes across the codes and merged the codes into workable categories. We then discussed the resulting categories along with their impact on the system design before creating a written documentation of each theme and their included subcategories. For better readability, we present a summarised report of our findings, aggregated by comments on different review types, students’ preferences, and requested changes.

4.2 Evaluation

When asked for their preferred review type, 41.58% of students well to the task content. This means that answering such questions was seen as particularly problematic as the questions were mandatory, and not answering a question or even just referring to a previously answered question was seen as not completing a task adequately. Other students mentioned that some questions within a review were too similar to each other, so that it felt like answering the same question again and again, adding little to the overall review. This was seen as particularly problematic as the questions were mandatory, and not answering a question or even just referring to a previously answered question was seen as not completing a review properly, which negatively influenced the assessment of their reviews.

4.2.1 Guided Reviews. Even though guided reviews were most preferred by students, there are still aspects that were criticised. The main critique by participating students regarded the wording of review questions. Students argued that some questions did not fit well to the task content. This means that answering such questions about the work at hand was hard or impossible. They also noticed that some questions were being re-used for multiple tasks, thus corresponding with neither task adequately.

Other students mentioned that some questions within a review were too similar to each other, so that it felt like answering the same question again and again, adding little to the overall review. This was seen as particularly problematic as the questions were mandatory, and not answering a question or even just referring to a previously answered question was seen as not completing a review properly, which negatively influenced the assessment of their reviews.

4.2.2 Open Reviews. The main criticism of open reviews was the lack of text formatting tools, as students only had a long text field to input their reviews. This was seen as a problem that affected both reading and writing reviews. Students reported losing their train of thought while writing what they perceived as a “block of text”. Recipients of reviews complained about a lack of structure, describing the texts as hard to read. This hampered the distinction between actionable feedback and miscellaneous side remarks.

This also makes it harder to map feedback from the review to the corresponding parts of the reviewed work to the point of being impossible. Some authors were left to wonder as to what possible improvements were proposed to them, and how to react to the given feedback.

4.2.3 Checkbox Reviews. Students mentioned that the main advantage of checkbox reviews was that they were short and easy to get done. Especially students who did not accept reviewing as a learning and reflection opportunity and suggested we should eliminate reviewing entirely named checkbox reviews as their favourite review type.

Students who actively disliked checkbox reviews characterised them as monotonous and boring, and failed to see the benefit of writing or receiving them. They mentioned that having to check the same or similar boxes for each task was a rather mind-numbing experience. They also complained that some checkboxes could not unambiguously be checked or respectively not checked. Students wished for a finer granularity of categories rather than only (rather) yes/(rather) no options. In other cases, the categories used were described as unclear and in need for a better explanation.

Finally, when receiving checkbox reviews, authors had troubles to understand why some checkboxes could be ticked without reviewers giving an explicit explanation. For example, it was possible to convey that a work was at least partially plagiarised without any further explanation. While reviewers had the option to explain their verdict in a separate text field, this was skipped often enough as it was not mandatory.

4.2.4 How to support and help students to write better reviews. In the survey, we posed several questions targeting possible improvements of the reviewing process to support students in writing better reviews, e.g. “How can we encourage students to write better reviews?” “How can we support you to write better reviews?”, or “Are you missing any functionalities while writing reviews?”. These answers were categorised as evaluation (63 mentions), structure (62 mentions), feedback (27 mentions), on-boarding (23 mentions), motivation (11 mentions) and impact (5 mentions).

Evaluation of reviews. Most students argued that the impact of written review quality on their grade should be considerably higher. A lot of time and effort goes into reviewing, which in their view is not adequately reflected in the grading. Students suggested upgrading the value of written review quality to become a proportional factor in the assessment. They argued that the current regulation failed to be an incentive to raise the quality over the basic threshold necessary. Others proposed that exceptional reviews, or alternately a higher accumulative quality rating at the end of the semester, should yield bonus points towards the grade. They proposed stricter or harsher evaluations of reviews, so that their peers stick to the reviewing instructions more closely, and/or more transparent evaluation scales to make adhering to the evaluation criteria easier.

A small number of students suggested we should evaluate reviews more lenient, or not at all. These students suggested this would reduce pressure, leading to more honest reviews. Students who preferred no review evaluation still suggested rewarding good reviews with additional points towards the final grade.

Structure. Many students suggested that we should reduce the quantity or scope of reviews to write, or at least that we correlate the length of reviews to write with the length of the reviewed work. Some would prefer no specifications at all for reviewing, while others who struggle with the lack of a specified minimum length of a review wish for more detailed specifications. Several students expressed doubt that the proposed feedback sandwich was helpful, arguing that forcing positive feedback does not always reflect the quality of the work, and consequently can feel insincere or made up. They also stated that receiving what felt like forced positive feedback did not help them in accepting criticism.
Feedback. Students remarked that receiving quicker and more comprehensive feedback on their written reviews from the course team would help them develop their review writing skills. They criticised the current feedback process regarding reviewing, only receiving textual feedback at the very end of a challenge, as being too late and not helpful for improving their reviewing skills. Additionally, they argued that the type of feedback they received could be improved, more accurately pointing out what they were doing well and what they could do better. Especially for average and bad reviews specific clues towards improvement would be necessary. For example, instead of “this review is too short”, they would need a description of what was missing in the review to better understand how to improve the quality of a review, as length alone is not an attribute of quality. In regard to feedback from their peers, students would like to communicate with their reviewers to clear up misconceptions.

Motivation. Students struggle with their own as well as their peers’ motivation to write reviews. Reasons for the lack of motivation are diverse: some do not perceive a positive influence on their, or their peers’ work due to reviewing; some find it takes too much time compared to a modest effect, which makes reviewing a waste of time and effort; some find reviewing too monotonous after some time, feeling that it forces them to be repetitive; others argue they would probably be more motivated if they in turn received better reviews from their peers.

On-boarding. Some students would profit from better on-boarding to the reviewing process. Students suggested we should offer more guidance for review writing in the form of lists, or even checklists of what to look out for, available while reviewing. A list of requirements for the current task would also help to identify what has been done and what should be improved. They also mentioned wanting more examples of good reviews to model their reviews after, and examples for bad reviews to see what pitfalls they should avoid. Examples for reviews should be available at the start of the semester, but should also be published for each new task as ongoing assistance. They wanted these new reviews to come from the current semester, and be published anonymously.

Impact. In addition to raising the impact of reviewing on the final grade, students suggested we should find ways to make quality reviewing valuable outside of the grade for this course. Proposals ranged from rewarding good reviews with badges; interconnecting this course with available soft-skill courses; reducing other requirements in our course; or disregarding their worst reviews if they wrote enough exceptional ones.

4.2.5 Other types of reviews students would like. As a final question we asked students what other types of reviews they would like to write or receive, resulting in a list of 18 other types of reviews. Some of these proposed review types were a direct answer to perceived issues with existing review types, such as using scales or drop-down menus in reviews rather than checkboxes. Also more visually categorised feedback was proposed such as spider diagrams or attribute based feedback as it is often seen for a character’s skill visualisation in video games.

Another proposition was to do comparative reviews, which was described as a review comparing a task to either the reviewer’s own work or other students’ work. While some questions in our guided review type already ask for just this type of comparison to the reviewer’s own work, students accentuated the advantage of seeing multiple examples of other students’ tasks simultaneously to gain a better understanding for a course “base line” of good or bad work, thus being able to adapt their feedback accordingly. They did, however, mention that this kind of reviewing would be more time consuming and as such would need to have a higher impact on their own grade.

A third type of review that was often mentioned can be summarised as using direct text manipulation. This was mentioned with two different goals in mind: some students wanted reviewing to be possible directly within the text, to highlight the context of the given feedback; others proposed a so called “text execution” review, in which reviewers could point out grammar errors, typos, wrong expression or similar issues with the work, which would require in-line edits to be usable.

Other new types of reviews included “subjective reviews”, where a review was completely comprised of review questions posed by the author about their own work; “argumentative reviews”, where a review was detailing the reviewer’s insights, or describing the best and worst of arguments raised by the author; “critical reviews”, where reviewers pose critical questions they had after reading the author’s work; or “creative reviews”, where reviewers build upon the author’s work by creating new ideas or follow-up arguments.

5 DISCUSSION

Our main goal for conducting this study was to find out how the current design impacts our students, and how it could be adapted to better fit their needs while simultaneously supporting the pedagogical and organisational grounding of the course. While students who filled out the survey were generally content with the current system, they also suggested many possible changes that might positively impact their work flow. Some suggestions indicate potential improvements, while others contradict goals and ideas manifest in the system.

5.1 Guided Reviews

The main criticism of guided reviews was the selection of questions students had to answer while reviewing. This critique might not seem like a usual design issue, however it greatly influences the students’ experience while reviewing, which is also discussed in related research [15, 21]. While we already put much work and research into composing review questions (briefly discussed in 3.1.1), this topic remains an ongoing issue that calls for specific research. Simply asking students to adhere to our review guidelines does not seem to yield the desired effects. Possible ways of dealing with this issue might be to communicate more transparently how and why review questions are chosen, e.g., if two review questions sound similar, we usually want students to highlight and explore different aspects of the author’s work.

The issue of questions not fitting to elaborations is an entire different challenge and might arise out of the general nature of the assignments our students are working on. Since most tasks involve discussions regarding wicked design questions, it is impossible to write review questions that fit to all elaborations equally, thus some questions are composed rather generally to enclose many different
types of possible elaborations. These might also be similar to such general questions in other tasks thus giving students the feeling of working on the same questions repeatedly. We do not see an immediate solution to avoid or reorganise these types of review questions in the current system.

5.2 Open Reviews

The introduction of open reviews can be ascribed to previous surveys conducted after students used the anonymous reviewing system, then with only guided reviews, for a semester. Students observed that they felt limited through review questions and that, if given the chance, they would highlight different aspects of their peers’ work, when being able to choose the structure of their review freely.

While the related work points towards textual reviews being more useful when structured by questions [15, 18] or by using triggers and prompts [21], students who answered the survey did not note any issues with formatting or struggling with content while writing open reviews. Nor did they mention their received reviews lacking depth. Getting the right message of the feedback across was ascribed mainly to formatting issues.

Rather than using long text boxes, which are sufficient for guided reviews as these already have a predefined structure, open reviews might be improved by allowing for more formatting options in the text fields. This will improve structure and readability, highlighting text, embedding links and adding resources. As a bonus, when including formatting options for reviews into the system, the same type of formatting options can also be offered to the authors for writing elaborations in the first place, thus providing a better workflow in two different aspects of the system.

A more difficult design issue, at least from a developer’s perspective, is the possibility to comment or mark elaborations directly in-line. Implementing such a change has to be seen as a longer-term plan.

5.3 Checkbox Reviews

We introduced checkbox reviews in answer to students complaining about the massive time investment they had to put into reviewing as it was found less time-intensive and easy to understand [11], thinking that a type of review that can quickly be completed might improve the acceptance among students. So it came as a surprise to see that students largely disliked this type of review and mentioned it the most when critiquing the reviewing system and when naming things that would profit from adaptation, change or discontinuation. A reason for this dislike might be that many students struggled with responsibly checking or not checking a box rather than having the opportunity to stay neutrally in the middle.

Checkbox reviews have the advantage that they provide a quick overview of all the requirements for a task as well as which have or respectively have not been met. This overview is not only usable by authors to see the main points of their work that might need a revision, but it is also an auxiliary overview for the course team when grading a challenge, as they can specifically look at the parts that are indicated as missing.

While we will not completely abandon this type of reviews, we plan to react to students’ suggestions by adapting possible question types to include more finely granulated sliders or scales instead of just checkboxes. On the other hand we are contemplating to introduce mandatory text fields that need to be filled out if a checkbox is not checked. Our concern regarding this change is, however, that students might be inclined to check a requirement as fulfilled rather than spend the extra time to fill out a reason why this might not be the case, again resulting in incorrect or unhelpful reviews.

5.4 Supporting students to write better reviews

Many suggestions pointed towards opportunities to facilitate and improve reviewing. Some of these can impact multiple of the aforementioned categories.

5.4.1 Evaluation of reviews. While some students argued that the evaluation of reviews led to higher pressure and less honest reviews, the experience of previous semesters has showed that some kind of evaluation was necessary to create and maintain a level of quality for reviews throughout each semester. To better reflect and highlight the effort that students put into reviews, we decided to change our evaluation approach. So far, all reviews written by the author of a work were evaluated while that work was evaluated, immediately yielding a finalised evaluation score for this author’s efforts. Henceforth, reviews will be evaluated together with the work they were written for. Consequently, an overall evaluation score can only be calculated when the work and all reviews have been evaluated separately.

This change comes with a big advantage for the course team. So far, when evaluating a challenge, we had to assess the reviews written by the challenge author. Thus, in order to appraise the quality of a review, we had to not only look at the students’ reviews, but also at the challenges they reviewed. This made the evaluation process longer and more complex. With the prospective change, we will evaluate the reviews that an author received for the work we are currently evaluating. While this means that students will have to wait longer for the overall evaluation of their work, this new workflow substantially reduces evaluation time for the course team, and at the same time provides for a more direct and transparent review evaluation for the students.

5.4.2 Structure. Asking several differentiated questions might make reviewing easier, as long as the number of questions corresponds with the length of the reviewed work. Usually, review questions are composed after the assignment is finished. They are aimed to cover as many qualities of a prospective elaboration for this assignment. The typical extent of such an elaboration was never regarded in the formulation of the review questions. So even if the assignment warrants an inquiry into multiple aspects of quality, the reviewed work might be too short to answer all these questions meaningfully, especially given that we ask for more than monosyllabic answers. A possible remedy would be to adjust the scope and length of review questions to the expected length of the elaboration to be reviewed. Given that writing textual feedback can impact motivation and distract from other tasks [3], the work load it creates should be carefully planned.

Reflecting recent criticism on the feedback sandwich [34, 43], we will examine alternatives to this form that place an emphasis on the intended effect - respectfulness - rather than the structural aspects.
5.4.3 Feedback. While restructuring the review evaluation by adding an extra evaluation field to each review, we can also add a feedback field that becomes mandatory as soon an evaluation score less than 100% is awarded. Keeping the already tight time frame allotted to each challenge evaluation by the course team in mind, such review feedback cannot take up copious amounts of time and effort. Short and concise feedback might still go a long way to help students improve and better understand how to hone their reviewing skills.

5.4.4 Motivation. Regarding the students’ motivation to write good reviews, there are no easy solutions. The context, a course mandatory for 750 first-year students, already provides for ample opportunity for some students to be disinterested in the course content, its requirements or structure. Additionally, most participants are first-year students, and many have no prior experiences with peer reviewing, neither writing nor receiving reviews. Consequently, they have no experiences in writing helpful reviews, nor do they know how to accept well-meant criticism, both aspects that need to be learned before they can be appreciated.

This is further complicated by the fact that we, as part of the course team and authors of tasks as well as review questions, are also involved in a continuous learning process of what works well and what does not work in composing such questions. While we rely on review question guidelines assembled from years of experience, it is still challenging to come up with new questions from scratch. We sometimes fall back on adapting questions written for previous worst practice examples, it still seems to be not enough to help students feel comfortable with the reviewing process. Many students called for more real-life example reviews. However, due to ethical and privacy issues, we decided to refrain from such a change: real-life worst practice examples could greatly impact the motivation or even well-being of students whose reviews were selected, even if these examples were to be published anonymously; real-life best practice examples on the other hand could negatively impact the students whose work was reviewed, as good reviews do not only praise but also criticise the author’s work. Instead, we can create more fictional examples for best and worst reviews and display praise but also criticise the author's work. Instead, we can create more fictional examples for best and worst reviews and display

5.4.5 On-boarding. While we already employ a whole challenge with the sole purpose of on-boarding that also includes best and worst practice examples, it still seems to be not enough to help students feel comfortable with the reviewing process. Many students called for more real-life example reviews. However, due to ethical and privacy issues, we decided to refrain from such a change: real-life worst practice examples could greatly impact the motivation or even well-being of students whose reviews were selected, even if these examples were to be published anonymously; real-life best practice examples on the other hand could negatively impact the students whose work was reviewed, as good reviews do not only praise but also criticise the author’s work. Instead, we can create more fictional examples for best and worst reviews and display them along with annotation, highlighting what can be learned from these examples.

Another potential strategy to create a better on-boarding experience for students would be to increase the difficulty of review questions throughout the semester. We plan for review questions that are easier to answer for the first few task and start to introduce more demanding review questions when students have gotten used to the process.

5.4.6 Impact. Some propositions regarding the possible impact of the efforts students expend reviewing outside the value with our course point to interesting opportunities. However, these do not pertain to this specific course and curriculum context and will thus not be discussed in the context of learnings we can draw from them.

5.5 Other review types

Students proposed a plethora of other review types. Our take away is to create a more modular and extended review system to allow for variations, while not relinquishing the simplicity of creating review types. This adaptation allows for some of the proposed new types to be tried out. Others can even more easily be implemented, since they mostly just require a change in how the questions are posed. Of course, such review types would have to be evaluated in their usefulness compared to textual feedback [22].

A few suggestions, such as the task comparison between two peer tasks that mirrors [9], would call for a larger change in system architecture and need to be excluded for now.

6 CONCLUSION

For more than ten years we have been working on the challenge to create a reviewing system that supports our students in their learning effort, while at the same time making the amount of work demanded for assessment manageable in a large university course. Our approach, utilising peer reviewing to reach this goal, has posed a set of specific problems regarding the motivation of students and the quality of their work. We have discussed the results of a survey to assess the introduction of different review types.

Our insights reflect findings from previous research, such as the questionable role of the review sandwich concept, or the value of good examples during on-boarding. They also raise issues specific to our approach like the rejection of our specific implementation of checkbox reviews. Our results pointed us to a crucial relation between the number of review questions and the length of a reviewed work, and ask for more research regarding specific strategies for the wording of review questions. These findings will guide the development of our course, our reviewing platform, as well as our future research.

ACKNOWLEDGMENTS

This work has been conducted in the context of the dCall project ‘Peer reviewing in large courses’, financed by the digital office of TU Wien. Much thanks goes out to our collaborators in this project: Michael Urbanek, Kay Kender, Lea Salome Brugger and Bernhard Kralofsky. Furthermore we would like to thank all students who participated in the survey and helped us better understand and design for their needs and motivations.

REFERENCES
