Int. J. Res. Undergrad. Math. Ed. (2023) 9:426–460
https://doi.org/10.1007/s40753-022-00167-y

# ConcepTests in Undergraduate Real Analysis: Comparing Peer Discussion and Instructional Explanation Settings

Thomas Bauer[1] · Rolf Biehler[2] · Elisa Lankeit[2]

Accepted: 19 February 2022 / Published online: 7 April 2022

## Abstract

Peer Instruction, first introduced by Eric Mazur in the late '90 s, is a method aiming at active student participation in lectures. It includes conceptual questions (so-called ConcepTests) presented to the students, who vote on answer alternatives presented to them and then discuss their answers in small groups. As professors have been reported to implement several variants of this method, it is highly desirable to understand the specific effects of the individual elements of the method (tasks, voting, and discussions in small groups). In the present study, we focus on the role of the discussion phase (peer discussion). Our study implemented two conditions: Peer Instruction in classical fashion, and a variant, in which peer discussion was replaced with instructional explanation by a tutor. Students in a course on Real Analysis were randomly assigned to the two conditions for two semesters. As far as learning outcomes are concerned, we do not measure these in terms of voting results within Peer Instruction cycles but we are focusing on transfer in terms of results in the final exams of the two semesters. Interestingly, we found no significant difference between the two conditions. Additionally, we had positive evaluations of the use of Peer Instruction in both variants, with no significant differences between the groups either. Regarding affective variables and learning strategies, no difference in the development could be detected. As an important practical implication, these results show that both implemented variants of the Peer Instruction method are justifiable as far as learning outcomes, measured by exam results, or students' assessment of the method are concerned. Our results put the widespread belief that it is mainly the peer discussion that accounts for the success of the use of ConcepTests into question.

**Keywords** Peer instruction · Analysis · ConcepTest · Peer discussion

✉ Thomas Bauer
tbauer@mathematik.uni-marburg.de

Extended author information available on the last page of the article

## Peer Instruction as an Instructional Strategy

Peer Instruction (PI) was introduced as an instructional strategy by Mazur (1996, 1997) in his introductory physics classes at Harvard university to foster active student participation and conceptual understanding. A cycle of PI consists of several elements: it starts with a conceptual question (a so-called ConcepTest), which concerns a relevant concept of the course. The question is presented to the students together with several answer alternatives. Students are requested to think about the question individually and commit to an answer in a first voting. Subsequently, the instructor asks the students "to try to convince their neighbors of their answer" (1997, p. 983) in peer discussions in small groups. After the phase of discussions (we refer to them as peer discussions henceforth), a second voting on the same question takes place, where students can make a new decision on one of the answer alternatives, based on the considerations during the peer discussion. Mazur reports that he consistently observes an increase in the proportion of correct answers in the second voting. Mazur's tasks are based on a conception of what constitutes conceptual understanding in physics, where overcoming mis- or preconceptions is a crucial element of learning. Thus, the idea of the peer discussion is not so much to create activity per se, but to foster conceptual understanding by targeting misconceptions (cf. Crouch & Mazur, 2001). Furthermore, the student activity and interactivity (in the sense of Chi, 2009) specifically emphasizes peer learning (Wentzel & Watkins, 2011). PI, and peer learning in general, has been shown to be an effective method of instruction, and one that students value (Balta et al., 2017). Since Mazur's first publications on PI (1996, 1997), the method has received much attention: Fagen et al. (2002) report about the widespread use of PI in physics, and also more generally in STEM education, such as in chemistry, life sciences, and engineering. In mathematics, PI has been deployed mainly in the entry phase of the university (e.g. Lucas, 2009; Pilzer, 2001). Terrell (2003) and Miller et al. (2006) show examples of using the method in a calculus class. Bauer (2019a) addresses the problem as to how PI questions can be constructed for a course on Real Analysis. The tasks he has developed for PI (Bauer, 2019b) are based on a theoretical model for task design based upon an extension of Tall and Vinner's (1981) construct of concept definition and concept image.

When it comes to concrete implementations, Turpen and Finkelstein (2009) report that physics instructors deploy various variants of PI when they implement the method in their courses. Therefore, both from a theoretical and a practical point of view, it is essential to understand the specific role that the various elements of PI (task design, voting, peer discussion) play, and to examine their specific effect on the reported positive outcomes. Our aim with the present study is to examine the role of peer discussion. Even though it is generally considered a central element of PI (e.g., Crouch et al., 2007), its precise role concerning learning gains is unclear. Our study compares PI (including a peer discussion phase) with a variant that replaces the peer discussion with instructional explanation by the tutor. The study was carried out in tutorial sessions during a course on Real

Analysis over two semesters. The students were randomly assigned to the two conditions (peer discussion vs. instructional explanation). Our goal in this study is to understand whether there is a difference between the two conditions concerning learning outcomes and students' perception of the use of the conceptual questions.

## Prior studies into Peer Instruction

### Effects of Peer Instruction on Learning Outcomes

(a) Why Peer Instruction May Work

We anticipate different reasons why PI would be beneficial to students' learning:

1. PI is a form of cooperative learning that aims at engaging students in discussions on the relevant concepts of the course. As such, it promises benefits for learners since there is a body of research that associates interaction with peers to advances in problem-solving skills, conceptual understanding, and metacognitive reasoning (see the review by Wentzel & Watkins, 2011) that takes place during the peer discussion phase. In addition, the fact that the discussion in a PI cycle is prompted through the presented answer alternatives enables focused information processing in the sense of Renkl and Atkinson (2007): the discussion is geared towards those aspects that the teacher (in their role as task designer) deemed most relevant.
2. PI makes use of audience response systems (ARS) – here, we use the term ARS in a broad sense, as the rounds of voting which are part of every PI cycle can be conducted by raising hands using voting cards. The initial voting in a round of PI thus constitutes testing, where the individual tries to retrieve and process remembered knowledge. The well-known "testing effect" describes the phenomenon that testing constitutes a learning event in and of itself, which can be even more effective than restudying. Even though not entirely understood theoretically, the effect has been well established under various experimental conditions (see e.g., the meta-analysis by Rowland, 2014). More significant effects are generally observed when testing is combined with feedback (i.e., with a representation of the initially studied information after a testing opportunity), but even without feedback, it has been shown to occur reliably (loc. cit., p. 15). DeLozier and Rhodes (2017), in their review on different approaches to the flipped classroom, point out that "the benefits of testing are robust and likely to enhance performance regardless of how it is carried out." In the same vein, Simelane and Skhosana (2012) report on the impact of the voting process itself. Indeed, they intentionally focus on the assessment and evaluation aspect rather than conceptual learning, using clicker questions as a means for evaluation and feedback every week. The element of testing itself seemed already to lead to increased involvement and participation of the students. As the voting element is not the focus of the present study per

se, we do not elaborate on this aspect. Instead, we refer the reader to Kay and LeSage (2009), who provide an overview of the possible benefits of using audience response systems.

3. In Mazur's implementation, PI is a constituent in a flipped-classroom instructional design, where the students are assigned pre-class readings. Thus, a substantial part of class time can be devoted to PI (Mazur, 1997, p. 983). When PI is part of such a substantial change in a learning organization, then beneficial effects may result from various elements in the setting – for instance, when pre-class reading leads to a deeper engagement of the students with the material. We refer the reader to DeLozier and Rhodes (2017) for a comprehensive literature review on flipped classrooms.

### (b) Learning Outcomes and Learning Gains in Peer Instruction

PI has been compared with traditionally taught courses both qualitatively and quantitatively. The studies differ regarding the type of PI implemented, regarding how learning gain was measured and how the comparative study was designed. If courses differ concerning PI and in other respects, it is not easy to attribute learning gain to the specific type of PI. Crouch and Mazur (2001), as well as Crouch et al. (2007) report increased mastery after introducing PI in physics courses concerning two measures: First, they find a substantial average normalized gain in the force concept inventory (FCI) from 1990 to 1991 (after introducing PI in 1991), and a slight improvement in a mechanics baseline test (MBT), which assesses quantitative problem-solving. In addition to the limitation that PI had been implemented together with the substantial change of implementing a flipped-classroom design, it is another limitation of these studies that the improvement is observed in courses in different years. It is conceivable, but not known from the literature, that the course itself might have changed over the years regarding content as well as regarding instruction principles to be more in line with the FCI.

In mathematics, PI has been studied mainly in Linear Algebra and Calculus. Pilzer (2001) reports on use of PI in a Calculus class. He observed a "significant improvement in reasoning skills" (Pilzer, 2001, p. 190), which was a relatively informal observation for the semester not backed up by data, and "impressive class retention" (p. 190) when ConcepTests were presented again at the end of the semester. For each ConcepTest, more than ninety percent of the students answered correctly. Additionally, a comparison of small cohorts (13 vs. 17 students) having been taught traditionally or with PI at the beginning of the next semester showed that average results of students taught via PI were higher both in conceptual and in standard problems in a short test covering material of the previous semester. In their Good Questions Project, Miller et al. (2006) show use of PI in a calculus course with 17 sections, where instructors were free in their choice of teaching method – including the decision whether or not to use PI. The instructors were surveyed on their use of PI, and in the final exam, students in groups whose instructors used PI scored significantly higher. Finally, Lucas (2009) collected data from his use of PI in a calculus course and found a positive correlation between students' average "I-clicker score" (accounting for participation and correct answers in the two rounds of voting) and their final course grade. He interprets these findings as PI enhancing students' comprehension. More recently, Cronhjort et al. (2018) compared lecture-based courses with courses using a flipped-classroom approach in eight study programs, four of which (assigned not by random but by instructors' preferences) used flipped classroom teaching. The students in the flipped classroom groups had higher learning gains. However, the authors do point out the limitation that different programs are compared in this study.

So, the beneficial effects of PI are widely reported. However, in many studies, PI is part of a flipped-classroom design. Therefore, what the specific contribution of PI is, as opposed to other constituents of the instructional design (such as pre-class reading assignments or a generally increased focus on conceptual understanding on the part of the teacher) is not clear. Researchers have also tried to gain insight into the learning gains that can be specifically attributed to the method of PI itself. For example, Rao and DiCarlo (2000) report an increase of correct answers between the two rounds of voting using PI within a medical physiology class. They interpret this increase as an indication of enhanced conceptual understanding (see also Cortright et al., 2005).

## The Role of Peer Discussion in Peer Instruction

When positive effects of PI are reported, e.g., in terms of voting results, it is unclear to which elements of this approach these effects can be ascribed. A first attempt to investigate the specific effects of the *peer discussion* was made in the study by Nicol and Boyle (2003), who compared two different discussion sequences: PI on the one hand and "class-wide discussion" on the other hand. Using semi-structured interviews and surveys, the study found that students perceived PI as more beneficial. However, a limitation of that study is that, although the two conditions were used in the same class, they were used on different topics and at different stages in the course. A comparison of perceived benefits is therefore tricky.

Smith et al. (2009) investigated the notion of measuring the effect of PI by the increase in correct answers between the two votes. They ask whether this increase could partly be due to *peer influence* during the discussion phase, i.e., stemming partly from the fact that knowledgeable students influence other students. Students influenced in this way may not necessarily gain conceptual understanding in the process. Their study, situated in an undergraduate introductory genetics course for biology majors, investigates whether the increased number of correct answers is retained in a subsequent vote on an "isomorphic" (loc. cit., p. 123) question. Concretely, they designed a modified PI sequence of the form.

$$\mathrm{Q1} \rightarrow \text{peer discussion} \rightarrow \mathrm{Q1}_{\mathrm{ad}} \rightarrow \mathrm{Q2}$$

where Q1 is the initial question posed in the PI cycle, which is posed again as $\mathrm{Q1}_{\mathrm{ad}}$ after the discussion. The new element is a question Q2, posed after $\mathrm{Q1}_{\mathrm{ad}}$, which is "isomorphic" to Q1 in the sense that only the "cover story" (loc.cit.) is changed, whereas it requires “application of the same principles or concepts for solution” (loc.cit.). Due to the increase in correct answers from Q1 to Q1$_{ad}$ and Q2, the authors conclude that most students learned from the discussion of Q1. However, it could be objected that a superficial transfer from Q1 to the isomorphic question could have occurred without a deep understanding of the correct answer to Q1 and its reasons. Porter et al. (2011) replicate this study design in upper-division computing courses (Computer Architecture, Theory of Computation). They found that results in subsequent isomorphic questions were not as high as in the study by Smith et al. The authors suggest that there might be differences between disciplines. In particular, the question as to when instructors identify questions as “isomorphic” might be answered quite differently between disciplines.

### Students’ Views on Peer Instruction

In addition to measured learning gains, how students perceive the use of ConcepTest is also of interest. Do they like using them? Do they feel they help understand the concepts, and do they find that the additional feedback increases their awareness of their knowledge level? How actively do they engage with fellow students and the tutor regarding the ConcepTests, and how helpful do the students rate this engagement for their understanding?

In their literature review on PI, Vickrey et al. (2015) state that “students have neutral to positive views on PI and seem to recognize its value over traditional teaching” (p. 4). Students often recommend using PI in future courses. Vickrey et al. (2015) conclude that students value different aspects of PI, especially the immediate feedback it provides, and, most importantly, they report that PI helps them learn the course material.

### Effects of Peer Instruction on Affective Variables and Learning Strategies

Some studies have investigated the effect of PI on students’ self-efficacy. For example, Zingaro (2014) found that computer science students taught using PI showed higher self-efficacy at the end of the semester than students taught traditionally in the same course by a different instructor. He attributes this to the “numerous opportunities for quick, accurate feedback in the PI class, where students could experience small successes” (Zingaro, 2014, p. 376). In addition, Gok (2012) compared students in algebra-based physics courses with or without PI and found that students’ self-efficacy in the treatment group increased significantly more than in the control group.

Gok (2012) also investigated other components of motivation. Regarding the value component, which includes students’ interest, and the affective component, which includes anxiety, no differences between students taught with and without PI were found.

Gok and Gok (2016) looked into the effects of PI on chemistry students’ learning strategies and resource management strategies. They found significant differences in the self-reported usage of these strategies at the end of the semester (but none at the beginning) between the group taught with PI and the traditionally taught control group: students taught with PI reported using the strategies of rehearsing, organization, elaboration, critical thinking, help-seeking, peer learning, metacognitive self-regulation, effort regulation and management of time and study environment to a much greater extent, with the most significant differences regarding peer learning and elaboration.

In summary, there is (a limited amount of) evidence for increased self-efficacy and increased use of learning strategies when students are taught with PI (compared with students taught without PI). Regarding other affective variables such as interest and anxiety, effect of PI is yet to be found. However, studies in this area are relatively scarce. There are especially no studies regarding the impact of the peer discussion phase on affective variables or learning strategies and no studies on mathematics students.

## Aim of this Study and Research Questions

Instructors using PI often report that it leads to increased involvement and participation of their students (e.g., Lucas, 2009). Also, some studies have reported that students perceive PI as fostering their understanding (see Vickrey et al.'s review on this matter). However, such self-reported benefits do not automatically allow conclusions about how much learning gain was achieved, as "students' perceptions of learning are not tantamount to objective measures of learning performance" (DeLozier & Rhodes, 2017, p.142). When it comes to objective learning outcomes, the results available in the literature so far are not clear. Several studies confirmed that generally more correct answers appear in the second voting on a PI question, and they consider this effect as an indication of increased understanding. This interpretation leads to the question: To what degree does this increase stem from peer influence (through knowledgeable students), and to what degree does it stem from actual peer learning? The study by Smith et al. (2009), carried out in a biology class, demonstrates positive effects on isomorphic questions posed immediately after a PI cycle. These effects are often taken as evidence that the effect of PI does go beyond peer influence. On the other hand, Porter et al. (2011) show that such results do not necessarily extend to other disciplines, where the concept of "isomorphic" questions might be viewed differently. Moreover, the question to what extent PI leads to transfer as measured by final exams is also open.

Studies such as Cronhjort et al. (2018) compare PI with traditional kinds of teaching in which ConcepTests are not used at all. When in such studies benefits are found within the intervention group (using PI), then these studies cannot distinguish between the effects stemming from the use of ConcepTests and the effects stemming from the specific way in which the ConcepTests are being used (peer discussion), since the two variables are confounded.

We aim at addressing the research gap described above in the following ways:

- Our study investigates PI not as part of a flipped-classroom design, but it focuses on the method of PI itself, which we understand as the combination

of voting on a multiple-choice question, subsequent peer discussion, and a second voting. In this sense, we view PI as a method that can be used as an intervention in selected parts of a course, independently of an overall design decision concerning a flipped-classroom setting.
- The study focuses on peer discussion, i.e., the phase during a PI cycle, where students discuss the solutions on offer to a multiple-choice question to convince the other group members. As peer discussion is considered an essential part of PI (Mazur, 1997; Pilzer, 2001), we aim at investigating its specific effect by comparing it to tutorial explanation leaving all other elements of PI the same.

We focus on learning outcomes as measured by end of semester exam results that are aligned with conceptual understanding as required in the ConcepTests, rather than assessing understanding directly after the interventions.

In order to achieve these aims, our study compares two groups of students who attended the same course (for two semesters), which consists of weekly 4 h lectures by the first author and 2 h exercise sessions run by experienced graduate tutors. The students were assigned randomly to two groups:

- In group A, PI was used in the exercise sessions every week. In this group, ConcepTests were used as suggested by Mazur, including a phase of peer discussion with each ConcepTest and second voting after the peer discussion.
- In group B the same ConcepTests were used. After being presented with a ConcepTest, students had time to consider the question individually, and they were then asked to vote for an answer. Only this single voting took place, and the tutor subsequently explained the solution (with limited contributions from individual students).

The precise instructions for the tutors' behavior in each group are described below.

To address the research gap identified above, our setting has two characteristics:

- We investigate the benefit of peer discussion vs. individual thinking combined with instructional explanations when the same tasks are being used. The difference thus does not lie in the mathematical content (and in the task design) but solely in the instructional procedure.
- We use exam results as a measure of learning outcome to detect possible effects that go beyond immediate effects on isomorphic questions.
- We ask students of both groups how they perceive the use of ConcepTests and how they perceive their activity in the exercise sessions to check for other effects besides learning gains.

We investigate the following research questions:

(RQ1) Is there a difference between the two groups in terms of learning outcome, as measured by the results in their final exams?

(RQ2) How do the students perceive the use of ConcepTests in tutorial settings concerning engagement? Is there a difference between the two groups in how students perceive the use of ConcepTests in their exercise sessions?
(RQ3) How do the students evaluate the implementation of the ConcepTests, and what changes do they suggest?
(RQ4) Is there a difference between the two groups concerning the development of affective variables and learning strategies?

Concerning RQ1, we had no clear hypothesis. Successful learning could occur in discussions with peers and through active listening to high-quality explanations from a tutor.

In RQ2, we were interested in finding out whether our different designs lead to different evaluations by the students regarding the benefits of ConcepTests. We try to determine whether the positive evaluations described above can be attributed to the peer discussion part or the general use of ConcepTests with voting.

We expected students from group A to engage more actively with fellow students than students from group B. Regarding the other question within RQ2, both directions were likely: that students from group A would perceive the ConcepTests as helpful for their understanding because of their active engagement and find that it gives them better feedback since they discuss with their peers and voice their explanations; and, that thinking about and voting on the ConcepTests in combination with the detailed explanations of the tutor improve students' understanding even more. Furthermore, the feedback the students receive by voting and hearing detailed explanations why their answer was right or wrong from an expert instead of their peers might also be even more helpful. It can also be assumed that students in group A enjoy using the ConcepTests more than students from group B because they interact with their peers, adding a social component to the activity.

Since, in our study, ConcepTests were not used in the lecture but in tutorial group meetings, "breaking up the lecture" or "opportunity to participate in the lecture" does not apply to our design. Therefore, whether the ConcepTests would be perceived in a similar way as described in the literature for their use in the context of lectures remains an open question.

Regarding RQ3, assessing students' perception of ConcepTests includes how satisfied students were with the concrete implementation and whether they had suggestions for future improvements. For instance, we wondered whether students from group A would wish for more detailed explanations from the tutor or group B students would wish for time to discuss the questions with peers. Questions concerning future improvements include whether more or fewer ConcepTests were desired and about the potential benefit of anonymous voting.

Regarding RQ4, we had no clear hypothesis. Improvement of students' self-efficacy or self-concept after using PI has been pointed out in a few studies, but the studies in question were not carried out in mathematics courses, and they did not contrast the use of ConcepTests with and without peer discussion.

On the one hand, there might not be any differences between the two groups because the effect of PI on affective variables and learning strategies might not stem from the peer discussion phase, and both groups received similar feedback by testing themselves in all rounds of voting. On the other hand, the assumption that mathematical self-efficacy and mathematical self-concept would change more in group A because their strengths and weaknesses might show more distinctly in discussion with their peers could prove plausible. Another assumption could be that in group A, students' enjoyment of math-related activities might increase substantially because of having fun in the peer discussions, which could seem more enjoyable than listening to the tutor's explanation. To successfully work with ConcepTests, specific learning strategies are helpful, e.g. linking and using examples. Gok and Gok (2016) showed the effects of PI on students' learning strategies in Chemistry. Therefore, one could think that students' usage of these would increase more in group A because, in this group, students needed to communicate their arguments and strategies actively to each other.

## Methods: Instruments, Design of the Study, Sample

### General Setting

Our study took place in the Analysis I course at the university of Marburg in the summer term of 2018 and the consecutive Analysis II course in the following winter term, both taught by the first author. These Analysis courses in Germany are more similar to a Real Analysis class rather than a Calculus class. Students enrolled on the Mathematics, Business Mathematics and Physics courses, as well as the pre-service mathematics teachers attend the Analysis I course, usually in their second semester. They can also take the course in a later semester, for example if they have to repeat classes they did not pass on the first try. At the university of Marburg, students can also start their studies in the summer term. If they do so, they are also allowed to take Analysis I in their first semester. We note that, in many universities in Germany, students are typically expected to take Analysis I in their first semester.

The Analysis courses we focus on in this paper comprised two 90 min lectures per week and one 90 min tutorial group meeting. Student tutors led eight different tutorial groups. Usually, in these group meetings, tutors present and discuss solutions to the weekly exercises and sometimes work on new tasks together. In these particular Analysis courses, the first thirty minutes of each tutorial group meeting were used for working with the designed ConcepTests (Bauer, 2019b) regarding the current lecture topics. We call these first thirty minutes "ConcepTests phase". In four groups, the ConcepTests were used in the specific way that Mazur proposed for PI, with a significant emphasis on discussions in small groups. We mentally pool these four groups and call them "Group A". More precisely, the procedure was as follows: First, the ConcepTest was shown to the students. They had 1–2 min for individual considerations without talking to each other and then voted by showing hand cards. The tutor then grouped the students into groups of two or three students who did not all vote for the same answer and asked them to convince each other of their choice. For this peer discussion, the students were given five minutes. Afterwards, second voting with hand cards took place. Then, the tutor announced the correct answer and explained briefly why this answer was correct. They did not comment on the wrong answers. Questions from the students were permitted and were answered by the tutor.

The other four groups (pooled as “Group B”) used the same voting questions but worked with them differently: the students voted once on the ConcepTests after individual consideration (1–2 min). Then, one student who answered correctly could volunteer to explain their reasoning. Afterward, the tutor gave an in-depth explanation of the correct solution. In this explanation, the tutor explained why the correct answer was correct and why the others were wrong and which common misconceptions these answers were based on. Therefore, the focus in group B lay on instructional explanation, as opposed to the peer discussion phase in group A. Questions from students were permitted and answered by the tutor, but there was no peer discussion.

In both groups, the same total amount of time was allotted for the tasks: the tutors were asked to spend ten minutes on each ConcepTest (including showing the question and explanations) and thus 30 min in total for the ConcepTests phase. In addition, before the beginning of the semester, the tutors were trained to perform their respective variants and were given examples illustrating the desired depth of the explanations.

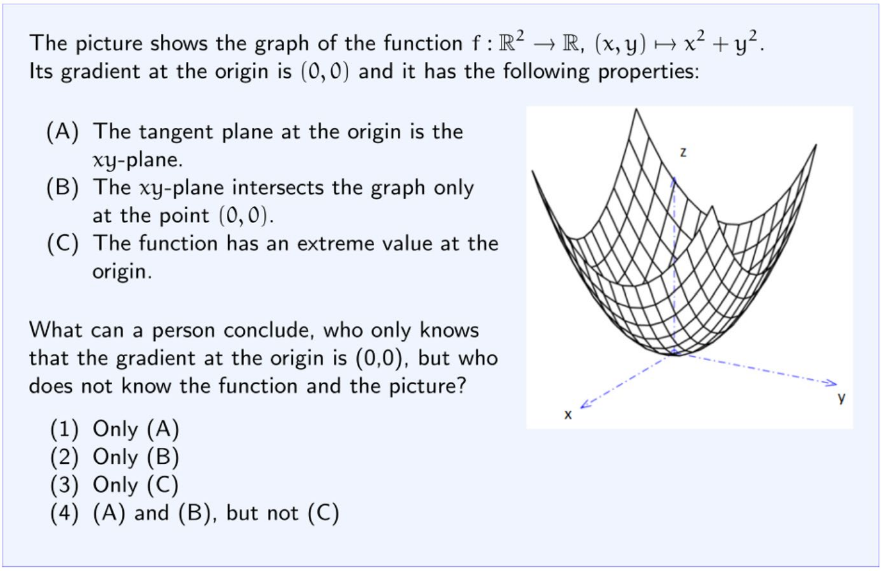
The random assignments of students were carried out very carefully as follows. There were four time slots for group meetings during the week, and in each slot, two groups met simultaneously (in different rooms). At the beginning of the semester, students chose a time slot for their tutor group meeting, but the particular group was assigned randomly. Each tutor group was assigned to be part of either group A or group B. At all times, one of the parallel groups was from group A and one from group B. Therefore, the students were assigned to group A or B randomly. Using the answers from the survey at the beginning of the semester, we additionally used the non-parametric Mann–Whitney-U-test to find out whether there were differences between the two groups regarding age, semester, study program, results in the earlier exams in their studies, affective constructs, or used learning strategies at the beginning of the semester. This was not the case. We, therefore, conclude that the two groups are comparable. Students were assigned randomly to group A and B in Analysis II again, so that there is no correlation between their group in Analysis I and Analysis II.

We conducted surveys via pen-and-paper at the beginning of the semester ($n = 125$) and the end of the semester once in the lecture ($n = 79$) and once in the tutorial group meetings ($n = 60$), each marked with a personal code for each student so that we were able to match the results from the different surveys with each other.

To address the first research question, we collected the students’ results in the end-of-term exam with their personal code, so that connecting the results with the earlier surveys was possible ($n = 89$). Moreover, we obtained their score for every single exam task. Additionally, we let them check a box on their exam which tutor group they attended during the semester.

## Design of the ConcepTests

The ConcepTests used in the tutorial sessions were designed according to the principles laid out in Bauer (2019a). In the first step of the design process, the central concepts of the material that was addressed in the current week's lecture were identified. Next, potential misconceptions about the concepts or theorems were identified. In this step, we use Tall and Vinner's (1981) concept image / concept definition to classify misconceptions. These misconceptions might be related to the concept definition (e.g., to using a precise definition in order to distinguish a notion from a related one) or to the concept image (being able to make sense of the concept, also in its relationships with other concepts, and being able to use the concept within the mathematical framework as well as in practical situations). Following the recommendation of Crouch et al. (2007), each ConcepTest thus focuses on an individual concept. For example, Fig. 1 shows a sample ConcepTest addressing the concept of *gradient* (see Bauer et al., 2022, in press, for a discussion of further examples). It requires an understanding of the relationship of the gradient with tangent planes and extreme values of functions of two variables to see that answer (1) is correct: only condition (A) can be deduced from the fact that the gradient is (0,0), even though, in the concrete example, also (B) and (C) are satisfied.



The picture shows the graph of the function $f : \mathbb{R}^2 \to \mathbb{R}$, $(x, y) \mapsto x^2 + y^2$. Its gradient at the origin is $(0, 0)$ and it has the following properties:

(A) The tangent plane at the origin is the $xy$-plane.
(B) The $xy$-plane intersects the graph only at the point $(0, 0)$.
(C) The function has an extreme value at the origin.

What can a person conclude, who only knows that the gradient at the origin is (0,0), but who does not know the function and the picture?

(1) Only (A)
(2) Only (B)
(3) Only (C)
(4) (A) and (B), but not (C)

**Fig. 1** A ConcepTest used in the section on differentiable functions of several variables. Translated from German

**Table 1** Overview of tasks in the final course exams. Only the tasks set in italics contained procedural components; all other tasks were based on conceptual knowledge

| **Final Course Exam "Analysis I"** | | **Final Course Exam "Analysis II"** | |
|---|---|---|---|
| 1 | [11 subtasks related to ConcepTests] | 1 | [11 subtasks related to ConcepTests] |
| 2 | Supremum | 2 | Metric spaces |
| 3 | Convergence of sequences | 3 | Differentiability |
| 4 | *Convergence of series* | 4 | *Extreme values* |
| 5 | *Limits of functions* | 5 | *Implicit function theorem* |
| 6 | Continuity | 6 | Integration |
| 7 | Differentiability | 7 | Jordan-measurable sets |
| 8 | Sequences of functions | 8 | *Differential equations* |

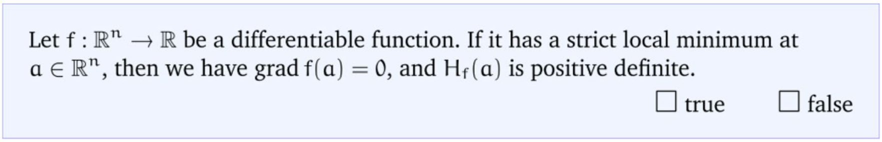
## Measuring Students' Learning Outcomes in the Final Exam

In the final course exams, a strong emphasis was put on conceptual questions: The exams in Analysis I and II consisted of eight tasks, only two (resp. three) of which had procedural components (Table 1 provides an overview of the exam tasks, and Fig. 2 shows a subtask related to the content of the ConcepTest from Fig. 1).

Task 1 of the exam was explicitly aligned with the ConcepTests used during the semester, both in format and in content: It consisted of 11 single-choice items, each of which was related to a ConcepTest that had been used in a tutorial session. Figure 2 shows an example of a subtask (related to the ConcepTest in Fig. 1). As we intended to avoid results stemming simply from remembered answers, the exam question is not identical to the ConcepTest, but instead relies on connections between the relevant concept images.

## Items and Scales Related to Students' Perception of ConcepTests (RQ2)

To answer our second research question, we developed items and scales concerning students' assessment of the ConcepTests and their use in the tutorial group meetings. An overview of these scales can be found in Table 2. Some of our items were based on stand-alone items used in questionnaires by Duncan (2005), Wolf



Let $f : \mathbb{R}^n \to \mathbb{R}$ be a differentiable function. If it has a strict local minimum at $a \in \mathbb{R}^n$, then we have $\text{grad}\, f(a) = 0$, and $H_f(a)$ is positive definite.

□ true □ false

**Fig. 2** A sub-task of task 1 in the exam "Analysis I", whose content is related to the ConcepTest shown in Fig. 1. Translated from German

**Table 2** Scales for students ' assessment of the ConcepTests and their use in the tutorial group meetings

| Name of scale *(number of items)* | Explanation | Example Item | Cronbach's Alpha |
|---|---|---|---|
| Self-report: Support of understanding of content by using ConcepTests *(3)* | Students report whether they think that the use of ConcepTests helped them understand the content more thoroughly | "The ConcepTests and their discussion helped me better understand the topic in question." | 0.902 |
| Self-report: Increasing awareness for own level of knowledge by using ConcepTests *(4)* | Students rate to what extent the ConcepTests contributed to their self-assessment of their knowledge, e.g. by making them aware of their mental images or their misconceptions or gaps of knowledge | "In the ConcepTest phase, you notice which misconceptions you have about certain concepts." | 0.809 |
| Personal attitude and emotions concerning ConcepTests *(5)* | Students report how much they like using ConcepTests, regarding enjoyment, fun, and interest and give a general evaluation of reasonableness | "I would rather attend a tutorial group meeting with ConcepTests than without them." | 0.924 |
| The intensity of active engagement with fellow students in the context of ConcepTests *(3)* | Students report to what extent they are actively engaged with fellow students regarding the ConcepTests, e.g., by explaining their answers, thinking critically about fellow students' explanations, or participating in discussions | "In the ConcepTest-phase, I explained why I chose my answer." | 0.843 |
| Perceived benefit of active engagement with fellow students for understanding *(3)* | Students report how helpful they found the activities mentioned in the scale above | "How helpful was this for your understanding?" [referring to the item above.] | 0.837 |
| The intensity of active engagement with the tutor in the ConcepTests phase *(2)* | Students report to what extent they actively engaged with the tutor regarding the ConcepTests, which can be asking questions to the tutor or actively thinking about his or her explanations | "I actively engaged with the tutor's explanations." | 0.639 |
| Perceived benefit of active engagement with the tutor for understanding *(2)* | Students report how helpful they found the activities mentioned in the scale above | "How helpful was this for your understanding?" [referring to the item above.] | 0.758 |

et al. (2014), or Butchart et al. (2009). We grouped them into scales using factor analysis together with content-based considerations.

Levels of reliability of all these scales are good or very good, except for the scale "Intensity of active engagement with the tutor in the ConcepTests phase" with Cronbach's alpha of 0.639. We decided to use this scale nevertheless rather than analyzing the two items separately. In addition, the scale "Perceived benefit of active engagement with the tutor for understanding" building on this scale shows better reliability.

### Items for the Evaluation of the Concrete Implementation (RQ3)

In addition to the scales shown in Table 2 that ask for the students' general assessment of using ConcepTests and their engagement in the ConcepTest-phase, we formulated stand-alone items for the evaluation of the specific implementation to answer RQ3. We asked the students to rate the average level of difficulty of the ConcepTests, the time for asking questions to the tutor, tutor explanations, time for discussion with fellow students, number of ConcepTests per meeting, and time for individual thinking before voting on a five-level Likert-scale (too little, rather too little, appropriate, rather too much, too much). Additionally, they were asked for requests for changes regarding anonymous voting, more discussion with fellow students, or more explanations from the tutor on six-level Likert scales. Details concerning the items will be reported in the results section.

### Items and Scales for Affective Variables and Learning Strategies (RQ4)

For the fourth research question, we used established scales regarding the affective variables mathematical self-efficacy (Ramm et al., 2006, adapted to the university context in the WiGeMath project (Hochmuth et al., 2018, Kuklinski et al., 2018)), mathematical self-concept (Liebendörfer et al., 2020, 2022, modified from Schöne et al., 2002), mathematics-related anxiety (Biehler et al., 2018, modified from Götz, 2004), enjoyment of mathematics-related activities (Ramm et al., 2006, adapted to the university context in the WiGeMath project (Hochmuth et al., 2018)) and interest in mathematics (Liebendörfer et al., 2020, 2022, inspired by Schiefele et al., 1993). An overview of these scales can be found in Table 3. The students additionally reported their use of learning strategies (Liebendörfer et al., 2020, 2022, see Table 4). All of these scales showed sufficient reliability (see Tables 3 and 4), except for the scale "Self-report: Use of the learning strategy 'Time investment'" with Cronbach's Alpha of 0.587 in the first survey. We decided to use the scale nevertheless because of the better values in the second survey. Removing single items did not improve the scale.

We analyzed the development of affective variables and the use of learning strategies only in the Analysis I course. We used an ANOVA (Analysis of Variance) with repeated measurements and the group affiliation as a between-subjects factor to determine whether the affective variables and learning strategies developed

**Table 3** Scales measuring affective variables

| Scale (number of items) | Example Item | Cronbach’s alpha ($t_1$ /$t_2$) |
|---|---|---|
| Mathematical self-efficacy (4)<br>(4-level scale) | “I am confident that I can perform well in homework and exams in mathematics.” | 0.750 / 0.779 |
| Mathematical self-concept (3)<br>(4-level scale) | “I understand mathematics mostly… [poorly / well].” | 0.716 / 0.732 |
| mathematics-related anxiety (3)<br>(6-level scale) | “When I think about mathematics in my studies, I am worried.” | 0.830 / 0.838 |
| Enjoyment of mathematics-related activities (6)<br>(6-level scale) | “I look forward to mathematics-related events (e.g., lectures, tutorial group meetings).” | 0.837 / 0.809 |
| Interest in mathematics (9)<br>(6-level scale) | “Dealing with mathematics is not exactly one of my favorite activities.” (reversed) | 0.767 / 0.811 |

**Table 4** Scales measuring the used learning strategies, 6-level scales in each case

| **Scale (number of items)** | **Example Item** | **Cronbach's alpha ($t_1$ /$t_2$)** |
|---|---|---|
| Self-report: Use of the learning strategy "Linking" (3) | "I try to understand how new content relates to what I have learned before." | 0.856 / 0.834 |
| Self-report: Use of the learning strategy "Simplifying" (3) | "I try to break down complicated relationships to a manageable level first." | 0.699 / 0.813 |
| Self-report: Use of the learning strategy "Using examples" (3) | "I create examples for theorems in order to understand the statement." | 0.816 / 0.706 |
| Self-report: Use of the learning strategy "Finding practical relevance" (3) | "With new content, I think about what it means in the real world." | 0.791 / 0.834 |
| Self-report: Use of the learning strategy "Structuring content" (3) | "For larger amounts of content, I prepare an outline that best reflects the structure of the material | 0.819 / 0.768 |
| Self-report: Use of the learning strategy "Time investment" (3) | "Before the exam, I take enough time to go through all the material." | 0.587 / 0.735 |
| Self-report: Use of the learning strategy "Tolerating frustration" (3) | "Even if I do not make any progress at all in learning, I keep trying until I get it right." | 0.838 / 0.801 |
| Self-report: Use of the learning strategy "Drilling" (3) | "I learn algorithms by performing the procedure over and over again." | 0.840 / 0.835 |
| Self-report: Use of the learning strategy "Memorising" (3) | "I try to retain the material better by repeating it several times." | 0.860 / 0.845 |
| Self-report: Use of the learning strategy "Learning with fellow students" (3) | "When I have a solution approach, I discuss it with fellow students." | 0.913 / 0.873 |

differently in the two groups. In this analysis, we only included students for whom data from the beginning and the end of the semester were available, leaving n = 64 students. We also excluded students in their first semester (n = 10) in this analysis because it is known that affective variables like mathematical self-concept change noticeably in the first semester, e.g. due to the big-fish-little-pond effect (Marsh, 1987; Rach, 2014) or the first-time phenomenon (Di Martino & Gregorio, 2018), which could have confounded our data. Therefore, this analysis includes only n = 54 students.

## Results

### RQ 1: Is there a Difference Between the Two Groups in Terms of Learning Outcome, As Measured by the Results in Their Final Exams?

In the “Analysis I” end-of-term exam, 48 students of group A and 39 students of group B took part. The scores in the different tasks in both groups (mean and standard deviation) can be found in Table 5. In addition to this table, we have displayed the ratio of the mean score to the maximal score to adjust for the different maximum scores of task 1 in Fig. 3.

Our central result of the Analysis I exam is that there are no significant mean differences between groups A and B neither in the total test nor in any of the tasks, not even on the 10-percent level. The direction of the differences points to group B (without discussion), with only one exception (task 2).

**Table 5** Scores in the Analysis-I exam, group A (n = 48) vs. group B (n = 39)

| Task | Topic of task | Max. possible points | $M_A$ $(SD_A)$ | $M_B$ $(SD_B)$ | $p$ | $d$ |
|---|---|---|---|---|---|---|
| 1 | Single-choice items aligned with ConcepTests | 11 | 3.90 (3.03) | 4.05 (2.87) | 0.773 | 0.05 |
| 2 | Supremum | 8 | 2.02 (2.40) | 1.94 (2.02) | 0.768 | -0.04 |
| 3 | Convergence of sequences | 8 | 1.50 (2.06) | 1.89 (2.07) | 0.354 | 0.19 |
| 4 | Convergence of series | 8 | 3.39 (3.35) | 3.85 (3.33) | 0.606 | 0.14 |
| 5 | Limits of functions | 8 | 3.01 (2.31) | 3.78 (2.65) | 0.189 | 0.31 |
| 6 | Continuity | 8 | 2.07 (2.43) | 2.18 (2.58) | 0.880 | 0.04 |
| 7 | Differentiability | 8 | 1.63 (1.97) | 2.15 (2.37) | 0.238 | 0.25 |
| 8 | Sequences of functions | 8 | 2.83 (3.60) | 3.67 (3.62) | 0.190 | 0.23 |
| total | | 67 | 20.34 (14.08) | 23.50 (14.09) | 0.209 | 0.22 |

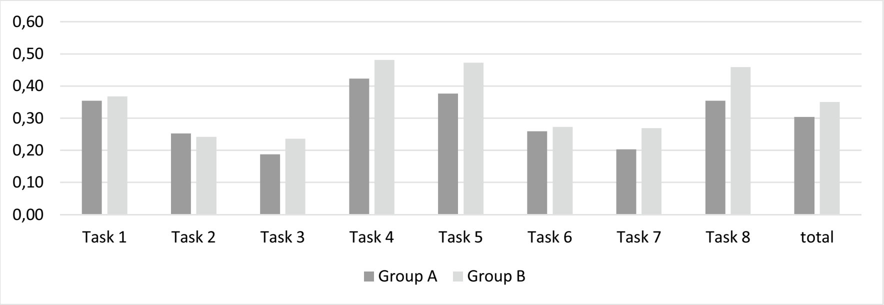


**Fig. 3** Normalized mean scores (mean scores divided by maximal possible scores) in the Analysis-I exam, group A (n = 48) vs. group B (n = 39)

In the Analysis II end-of-term exam, 32 students of group A and 24 students of group B took part. The results are displayed in Table 6 and Fig. 4.

Our central result of the Analysis II exam is that, again, there are no significant mean differences between groups A and B, neither in the total test nor any of the tasks, not even on the 10-percent level. However, the direction of the differences points to group A (with discussion), with only two exceptions (task 5 and 7) – this is a difference from the Analysis I results.

In the exam on Analysis I, a score of 27 (out of 67) was sufficient to pass; in Analysis II, a score of 32 (out of 66) was sufficient. The students were given a second examination opportunity within the same module. The overall pass rate in the module is in line with the long-term variability of such exams. Figure 4, compared to Fig. 3, shows much better results of the exam in this group of students, which may be a positive selection from the Analysis I students (attrition effect). We will not go into details as this is not relevant to our research question.

### RQ 2: How Do the Students Perceive the Use of ConcepTests in Tutorial Settings? Is There a Difference Between the Two Groups in How Students Perceive the Use of ConcepTests in Their Exercise Sessions?

Table 7 shows the results of the evaluation in Analysis I for all participants and each group. As the results in Analysis II were similar, we do not show them in detail here.

The results can be interpreted in two ways: The feedback concerning the use of ConcepTests as such and concerning differences between group A and group B.

**Group Differences** On a level of significance of 5%, the groups differ only in the intensity of active engagement with fellow students ($p = 0{,}029$, $d = 0{,}67$) and the perceived benefit thereof ($p = 0{,}045$, $d = 0{,}55$). Both of these means are higher in group A. The means of the other scales are also greater in group A, but these differences are not significant, not even on a level of significance of 10%. In other words, we do not find any difference beyond the most obvious and expected, namely the intensity of active engagement with students. This aspect was the design feature of group A.

**Table 6** Scores in the Analysis II exam, group A (n = 32) vs. group B (n = 24)

| Task | Topic of task | Max. possible points | $M_A$ $(SD_A)$ | $M_B$ $(SD_B)$ | *p* | *d* |
|---|---|---|---|---|---|---|
| 1 | Single-choice items aligned with ConcepTests | 11 | 8.03<br>(1.82) | 7.79<br>(1.98) | 0.822 | 0.13 |
| 2 | Metric spaces | 8 | 5.31<br>(1.60) | 4.83<br>(2.41) | 0.79 | 0.24 |
| 3 | Differentiability | 8 | 4.52<br>(2.61) | 4.31<br>(2.71) | 0.233 | 0.08 |
| 4 | Extrema | 8 | 5.39<br>(1.61) | 4.96<br>(1.59) | 0.284 | 0.27 |
| 5 | Implicit functions | 7 | 2.36<br>(2.36) | 2.94<br>(2.25) | 0.746 | -0.25 |
| 6 | Integration | 6 | 3.03<br>(2.10) | 2.88<br>(1.95) | 0.402 | 0.08 |
| 7 | Jordan-measure | 10 | 4.75<br>(2.96) | 5.31<br>(2.52) | 0.406 | -0.20 |
| 8 | Differential equations | 8 | 3.80<br>(2.30) | 3.33<br>(2.68) | 0.947 | 0.19 |
| total | | 66 | 37.19<br>(9.91) | 36.35<br>(12.46) | 0.797 | 0.08 |

The perceived benefit is consistent with other findings in the literature we quoted above.

**Whole Group** We find a very positive evaluation of the use of the ConcepTests as a whole: all the mean values are far beyond the theoretical mean of 3.5. For a detailed interpretation, we focus on the first three scales related to evaluating the method of ConcepTests. The students especially state having understood the Analysis concepts better by using ConcepTests (88% (rather) agree) and having gained awareness for their level of knowledge (95% (rather) agree). The whole distribution of answers can be found in the boxplots of Fig. 5. The first quartiles are each above the theoretical

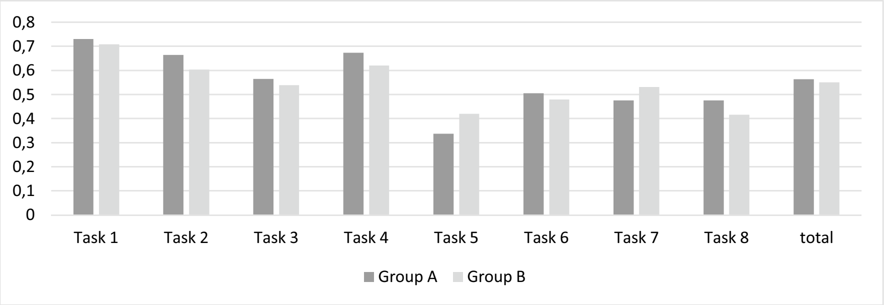


**Fig. 4** Normalized mean scores (mean scores divided by maximal possible scores) in the Analysis II exam, group A (n = 32) vs. group B (n = 24)

**Table 7** Evaluation of the use of ConcepTests in Analysis I. All Likert scales range from 1 (“completely disagree”) to 6 (“completely agree”)

| **Scale** | **Both groups (N = 58)** | **Group A (N = 27)** | **Group B (N = 25)** | | |
|---|---|---|---|---|---|
| | ***M (SD)*** | ***M (SD)*** | ***M (SD)*** | ***p*** | ***d*** |
| Self-report: Support of understanding of content by using ConcepTests | 4.79<br>(1.09) | 4.86<br>(1.15) | 4.68<br>(1.05) | 0.414 | -0.17 |
| Self-report: Increasing awareness for own level of knowledge by using ConcepTests | 4.79<br>(0.81) | 4.82<br>(0.88) | 4.66<br>(0.75) | 0.377 | -0.20 |
| Personal attitude and emotions concerning ConcepTests | 4.38<br>(1.34) | 4.37<br>(1.30) | 4.35<br>(1.29) | 0.927 | -0.02 |
| The intensity of active engagement with fellow students in the context of ConcepTests | 3.87<br>(1.39) | 4.28<br>(1.50) | 3.39<br>(1.13) | 0.029 | -0.69 |
| Perceived benefit of active engagement with fellow students for understanding | 4.18<br>(1.30) | 4.46<br>(1.24) | 3.77<br>(1.24) | 0.045 | -0.56 |
| The intensity of active engagement with the tutor in the ConcepTests phase | 4.11<br>(1.29) | 4.22<br>(1.30) | 4.02<br>(1.20) | 0.530 | -0.16 |
| Perceived benefit of active engagement with the tutor for understanding | 4.47<br>(1.27) | 4.50<br>(1.23) | 4.42<br>(1.27) | 0.802 | -0.07 |

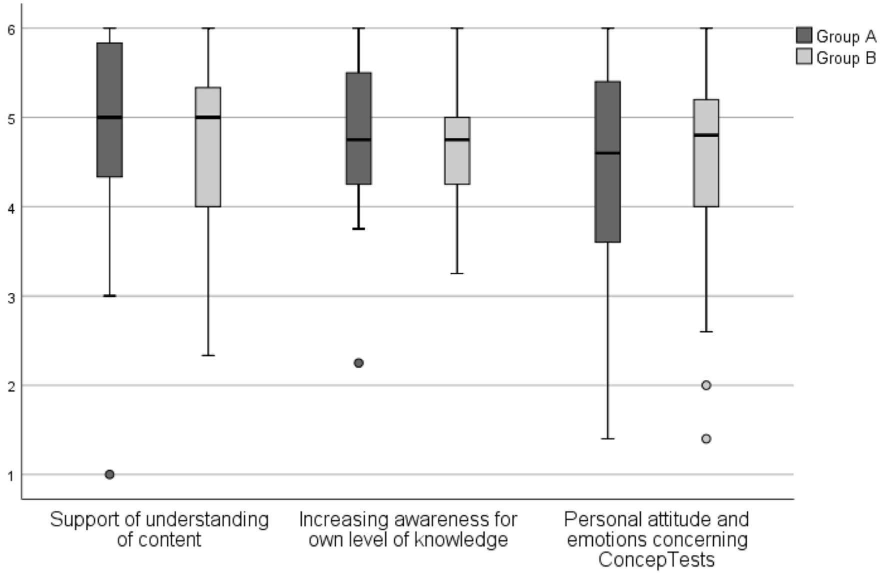


**Fig. 5** Distribution of answers to scales regarding students' view of ConcepTests in Analysis I, group A (n = 27) vs. group B (n = 25)

midpoint of the scales (3.5), meaning that 75% or more (rather) agreed. A more detailed summary of the approval ratings is shown in Table 8.

## RQ3: How Do the Students Evaluate the Concrete Implementation of the ConcepTests, and What Changes Do They Suggest?

We show again only results from Analysis I because the Analysis II results do not differ substantially.

In each group, a majority did not wish for more discussion or more explanations from the tutor. This result shows that each group was satisfied with their version without having experienced the other. Using a Mann–Whitney-U-Test, we found no significant differences between the groups, not even on a level of significance of 10%.

**Table 8** Percentage of students in each group agreeing to the scales regarding students' view of ConcepTests in Analysis I

| | **Percentage of agreement in group A (N = 27)** | **Percentage of agreement in group B (N = 25)** |
|---|---|---|
| Self-report: Support of understanding of content by using ConcepTests | 89% | 88% |
| Self-report: Increasing awareness for own level of knowledge by using ConcepTests | 96% | 92% |
| Personal attitude and emotions concerning ConcepTests | 92% | 88% |

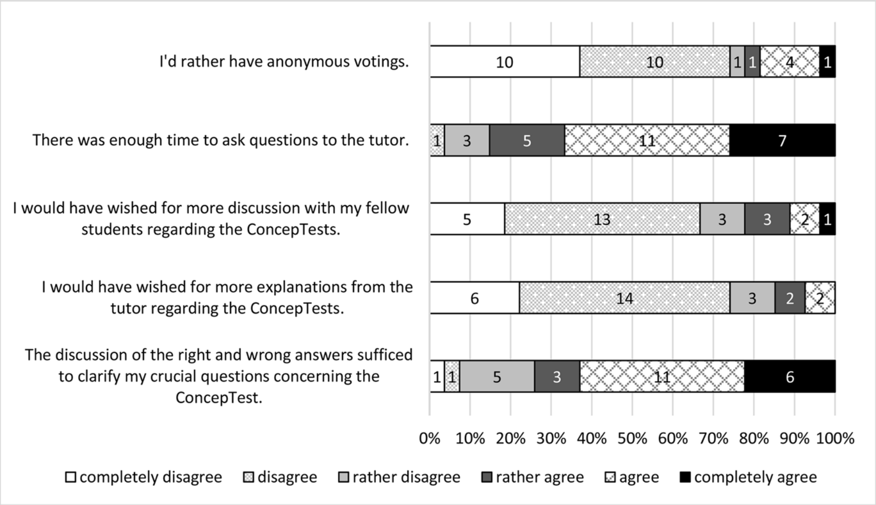


**Fig. 6** Evaluation and requests for changes, group A, Analysis I (n = 27)

Figures 6 and 7 show the results of the evaluations and requests for changes. About 25% in group A and 32% in group B would have preferred anonymous voting. Given that in group A, the voting behavior was decisive for the division into groups for discussion, it is not surprising that the proportion in group B was higher than in A. However, in both groups, only one person each expressed a strong wish for anonymous voting. Thus, the majority of the students were satisfied with the non-anonymous voting. The groups did not differ concerning any of the questions

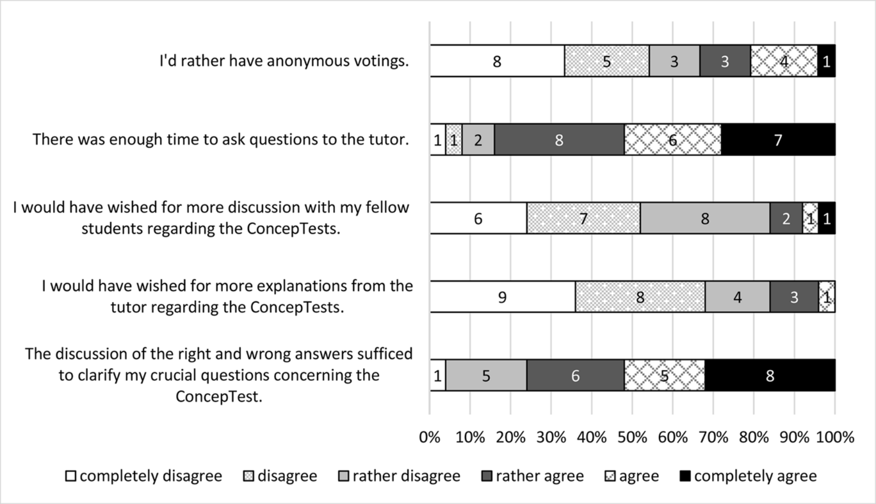


**Fig. 7** Evaluation and requests for changes, group B, Analysis I (n = 25)

(Mann–Whitney-U-test, level of significance 10%). Percentages of an agreement to the items about the ConcepTest phase can be found in Table 9.

Only for about 25%, the time was rather not sufficient to clarify understanding. Most of the students were satisfied with their version, wishing neither more time for discussion with fellow students nor more for explanations from the tutor or more time for asking questions. Remarkably, the percentage of students wishing for more discussion is not higher (and even lower, though the difference is not significant) in group B, even though they did not have a peer discussion phase at all. Additionally, the request for more tutor explanations is not higher in group A, which may be considered surprising as A's design contained only a short tutor explanation.

In addition to these requests for changes, the students rated the appropriateness of the average level of difficulty of the ConcepTests, time for asking questions to the tutor, discussion with students and individual thinking as well as the tutor's explanations and the number of ConcepTests per meeting (Figs. 8 and 9). The majority of students found all of this appropriate. Using a Mann–Whitney-U-Test, there was a significant difference in only one item: concerning the time for discussion with fellow students ($p = 0.015$). In group B, 40% of students rated it (rather) too little, while 70% of group A found it appropriate. Interestingly, as Figs. 4 and 5 show, this does not mean they would have preferred more time for discussion. From the design of group B, it is expected that this percentage is higher than in group A. Regarding time for tutor explanation, more students in group A rated it (rather) too little than in group B (not significant at the 10% level in a Mann–Whitney-U-Test). We expected a more considerable difference since the explanations in group B were much more detailed.

## RQ4: Development of Affective Variables and Self-reported Use of Learning Strategies

With a mixed ANOVA, we found no difference in the development of affective variables (mathematical self-efficacy, mathematical self-concept, math-related anxiety, enjoyment of math-related activities, and interest in mathematics) or learning strategies between the two groups from the beginning to the end of the semester, not even on a level of significance of 10%. We additionally tested for group differences at each point in time with a Mann–Whitney-U-test and found no differences on a level

**Table 9** Percentage of students in each group agreeing to the items regarding requests for changes in the ConcepTest phase

| Item | Percentage of agreement in group A (N = 27) | Percentage of agreement in group B (N = 25) |
|---|---|---|
| more time for questions | 15% | 16% |
| more fellow discussion | 22% | 16% |
| more tutor explanations | 15% | 16% |
| time not sufficient | 26% | 24% |

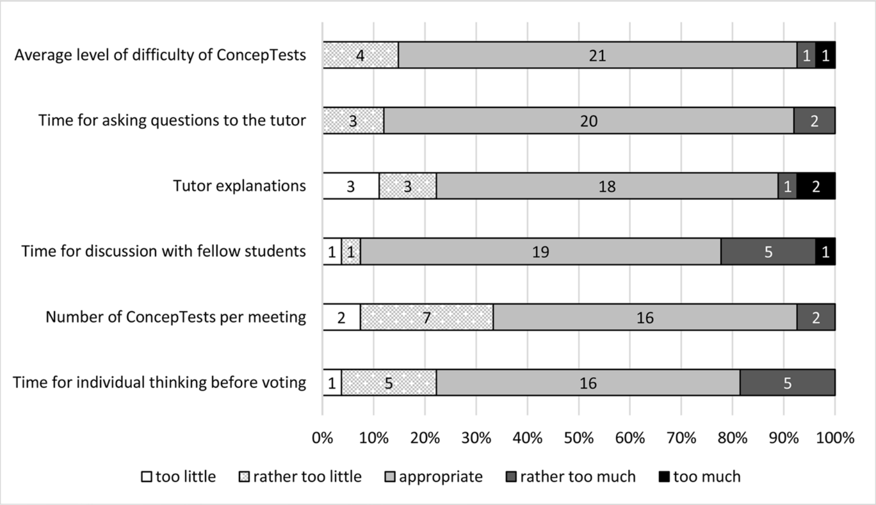


**Fig. 8** Students' evaluation of implementation, group A, Analysis I (n = 27)

of significance of 5%. For this reason and better readability, we show the results for the group as a whole together with $F$ and $p$ values from the ANOVA in Table 10.

Most of the variables we were interested in did not change significantly over the semester. However, we found that the students' mathematical self-efficacy declined ($d$ = -0.33, $p$ = 0.007) in the group as a whole. We also found that students reported using the learning strategy "linking" less ($d$ = -0.29, $p$ = 0.027) and the strategy "simplifying" more ($d$ = 0.31, $p$ = 0.011) at the end of the semester, compared to the beginning of the semester, while staying on a relatively high level.

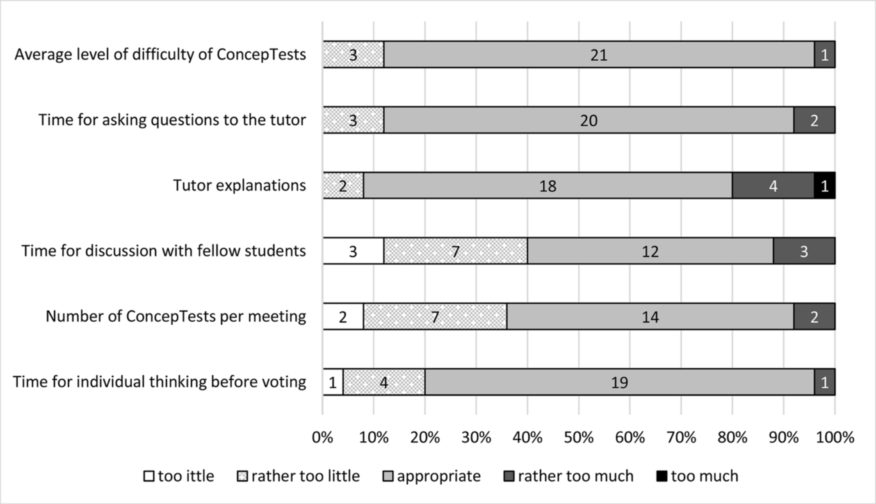


**Fig. 9** Students' evaluation of implementation, group B, Analysis I (n = 25)

**Table 10** Comparison of affective variables and use of learning strategies. [*$p < 0.05$, **$p < 0.01$ (Wilcoxon)]

| Scale<br>(from … to …) | Mean (SD) $t_1$ | Mean (SD) $t_2$ | Cohen's *d* | Results of mixed ANOVA: Interaction Construct*group ($t_1$ to $t_2$, A vs. B) | |
|---|---|---|---|---|---|
| | (whole group, without students in their first semester) | | | *F* | *p* |
| Mathematical self-efficacy<br>(1 – 4) | 2.91<br>(0.57) | 2.74<br>(0.47) | -0.33** | F(1,54) = 0.001 | 0.970 |
| Mathematical self-concept<br>(1 – 4) | 2.97<br>(0.49) | 2.94<br>(0.47) | -0.06 | F(1,52) = 0.441 | 0.510 |
| Mathematics-related anxiety<br>(1 – 6) | 3.41<br>(1.32) | 3.57<br>(1.28) | 0.13 | F(1,53) = 0.032 | 0.858 |
| Enjoyment of Math-related activities<br>(1 – 6) | 3.59<br>(0.86) | 3.52<br>(0.90) | -0.08 | F(1,52) = 1.667 | 0.202 |
| Interest in Mathematics<br>(1 – 6) | 3.63<br>(0.37) | 3.60<br>(0.37) | -0.09 | F(1,52) = 0.823 | 0.369 |
| Self-report: Use of the learning strategy "Linking"<br>(1 – 6) | 4.86<br>(0.68) | 4.64<br>(0.83) | -0.29* | F(1,53) = 0.421 | 0.519 |
| Self-report: Use of the learning strategy "Simplifying"<br>(1 – 6) | 4.60<br>(0.75) | 4.83<br>(0.74) | 0.31* | F(1,53) = 0.606 | 0.440 |
| Self-report: Use of the learning strategy "Using examples"<br>(1 – 6) | 4.89<br>(0.79) | 4.75<br>(0.91) | -0.16 | F(1,53) = 0.855 | 0.359 |
| Self-report: Use of the learning strategy "Finding practical relevance"<br>(1 – 6) | 3.23<br>(1.33) | 3.23<br>(1.32) | 0.00 | F(1,53) = 0.091 | 0.764 |
| Self-report: Use of the learning strategy "Structuring content"<br>(1 – 6) | 3.80<br>(1.25) | 3.90<br>(1.20) | 0.08 | F(1,53) = 0.124 | 0.726 |
| Self-report: Use of the learning strategy "Time investment"<br>(1 – 6) | 4.60<br>(0.90) | 4.60<br>(1.07) | 0.00 | F(1,53) = 0.395 | 0.577 |
| Self-report: Use of the learning strategy "Tolerating frustration"<br>(1 – 6) | 4.33<br>(1.11) | 4.45<br>(0.98) | 0.12 | F(1,53) = 0.483 | 0.880 |

**Table 10** (continued)

| Scale<br>(from … to …) | Mean (SD) $t_1$ | Mean (SD) $t_2$ | Cohen's *d* | Results of mixed ANOVA: Interaction Construct*group ($t_1$ to $t_2$, A vs. B) | |
|---|---|---|---|---|---|
| | (whole group, without students in their first semester) | | | *F* | *p* |
| Self-report: Use of the learning strategy "Drilling"<br>(1 – 6) | 4.22<br>(1.01) | 4.25<br>(1.00) | 0.03 | F(1,53) = 1.321 | 0.256 |
| Self-report: Use of the learning strategy "Memorising"<br>(1 – 6) | 4.45<br>(0.95) | 4.55<br>(0.94) | 0.11 | F(1,53) = 1.312 | 0.257 |
| Self-report: Use of the learning strategy "Learning with fellow students"<br>(1 – 6) | 4.84<br>(1.21) | 4.73<br>(1.30) | -0.09 | F(1,52) = 0.838 | 0.364 |

## Discussion of Findings, Limitations and Conclusion

We studied the specific role of peer discussion as part of a PI cycle during a two-semester course on Real Analysis ("Analysis I" and "Analysis II"). Group A in our study design used ConcepTests as suggested by Mazur, including a phase of peer discussion and second voting. Group B worked in a more teacher-centered setting, with time to individually consider the question and a subsequent explanation by the teacher. We aimed to compare the learning outcome, the students' perception of the use of ConcepTests, and the self-reported level of activity in the group meetings, development of affective variables, and learning strategies.

**Learning Outcomes** Our study measured the learning outcome through the grades in the final exams on Analysis I and Analysis II. Our findings show that none of the differences between groups A and B are significant at a significance level even of 10%. Interestingly, this applies to the total exam score and the results of any single exam task.

So, the different instructional use of ConcepTests (with peer discussion vs. instructional explanations) did not make a measurable difference in students' exam results between the two groups. We offer the following possible explanations. First, it is reasonable to assume that working with well-designed ConcepTests in and of itself can be of great benefit for conceptual understanding. In our study and previous studies, students reported that ConcepTests helped uncover misunderstandings and enriched their understanding of the concepts. The design of the ConcepTests may be more relevant than the specific way in which they are used. Note that in both instructional versions used in our study, the students engage in focused processing (in the sense of Renkl & Atkinson, 2007) when they consider the ConcepTest and commit to an answer. We certainly do not expect that in a design where ConcepTests would only be passively consumed, students could benefit in the same way. Instead, both study groups have benefited from the self-explanation effect (Fonseca & Chi, 2011). The ConcepTests here fulfill an activating role, much as DeLozier and Rhodes (2017) point out in their review on different approaches to the flipped classroom, where they even suggest that "the benefits of testing are robust and likely to enhance performance regardless of how it is carried out." Second, it is also well conceivable that the groups benefited from the respective instructional designs in different ways: While in group A the students had the opportunity to benefit from the advantages of peer learning (e.g., in discussing differing subjective conceptions, verifying or refuting arguments), the students in group B had the opportunity to benefit from instructional explanations by an expert, who might have been more able to provide relevant arguments and prototypical counter-examples than novices. The net effect could be similar to what Catambrone (1996) suggested (on a different research topic): It might be that "one advantage is matched by the other". Moreover, effects may be different for different student groups. One subset may exist where peer discussions are more effective and another subset where instructional explanations work better. If this were true, then one conceivable approach might be to always attach a phase of elaborated instructional explanation after the second vote. It is, however, an open question whether this could be detrimental to the PI scenario as a whole, as it might prove tedious for a considerable proportion of students. This is one of the reasons why further studies are needed to explore the effects on different subgroups of students.

**Students' Perception of the Use of ConcepTests** Groups A and B largely agree in their positive perception of the use of ConcepTests as far as their understanding of relevant concepts from the course is concerned as well as concerning increased awareness of their level of knowledge. The only significant difference is found in the intensity of communication with fellow students and its perceived benefit. The latter finding is as expected since communication in peer discussion is part of the study design in group A, whereas in group B the focus lies on explanation by the tutor. Interestingly, however, despite this difference in design, the students from neither group asked for more group discussion or more explanation from the tutor, being satisfied with their own version.

This result shows that both versions of using ConcepTests are evaluated positively by the students when they experience only one version. It is, however, conceivable that students who experienced both versions would prefer one over the other. However, we cannot conclude whether peer discussion or the tutor's detailed explanation would be more popular in direct comparison from our data. Also, it seems reasonable to expect that the preferences also depend on students' characteristics.

**Affective Variables and Use of Learning Strategies** Concerning affective variables, we found no different development in the two groups. Our hypothesis that mathematical self-efficacy and mathematical self-concept might change more in group A because their strengths and weaknesses might show more distinctly in discussion with their peers could not be confirmed. This observation could be explained since both groups voted on the same questions and received feedback about the correctness of their choice. Furthermore, the assumption that, in group A, students' enjoyment of math-related activities might increase more substantially because of enjoying the peer discussions could not be confirmed either. On the contrary, while not significantly different, the trend seemed to be that enjoyment decreased slightly in group A while staying on the same level in group B. This trend is only an indication and should not be overinterpreted.

We found a slight decrease in mathematical self-efficacy in the group as a whole. This result might seem surprising given that Gok (2012) found that students' self-efficacy taught with PI increased significantly more than in the control group. It is unfortunately common that students' self-efficacy decreases during an Analysis I course (e.g. Hochmuth et al., 2018). It is possible that students' self-efficacy decreased less in this course than in a regular Analysis I course taught without usage of ConcepTests, similar to the experiences regarding innovative lectures compared to regular lectures (Hochmuth et al., 2018). Due to the lack of a suitable control group, we cannot prove this hypothesis.

We did not find that the two groups used different learning strategies either. Therefore, the hypothesis that students' usage in group A might increase more because they had to communicate their arguments and strategies to each other did not show up in our data. This result indicates a similar use of learning strategies regardless of the occurrence of peer discussion.

Gok and Gok (2016) made a comparative study showing that students taught with PI made higher use of learning strategies and resource management strategies than the students in the control group. We did not compare to students with no ConcepTest phases but had the hypothesis that using learning strategies will increase after the intervention using ConcepTests and that the increase may be higher in the group with Peer Discussion.

However, we cannot confirm these hypotheses. The study by Gok and Gok (2016) was conducted with chemistry students, so one has to consider that there might exist disciplinary differences. Furthermore, students in our study reported using most of the learning strategies to a great extent already at the beginning of the semester, leaving not that much room for any increase. Our study found the following pre-post differences: students reported using the learning strategy "linking" less at the end of the semester than initially but using the strategy "simplifying" more frequently. However, we do not see how this is related to the ConcepTests; we would instead hypothesize that this might be due to the more advanced content at the end of the semester. Geisler (2020) also found a decrease in the use of the learning strategy "linking" among students in Analysis I (without using PI) during the semester, supporting our hypothesis that this is probably not linked to the ConcepTests.

The fact that we did not find differences in the development of affective variables or the use of learning strategies is an indication that students in both groups engaged similarly with the ConcepTests with and without peer discussion.

**Limitations of our Study** Firstly, the learning opportunities offered to the two groups differed only in the specific intervention for 30 min per week. On the one hand, this is a strength of our design, as it isolates the variable in question (i.e., the specific method of using ConcepTests). On the other hand, it is also a limitation, since one can argue that the difference between the two groups, who were in very similar learning environments over a long period of time, is too small to be observed at all. In view of our results, it would be very interesting (while probably ethically challenging) to design a study that, in addition to our groups A and B, includes a third group C, in which ConcepTests are not used at all. This could provide an opportunity to test the hypothesis indicated above that it is the use of ConcepTests per se that makes a difference (between A/B versus C) rather than the specific way in which they are used (A versus B). It would also be interesting to develop study designs that can serve to further explore the question as to what degree the learning effect comes from the students being required to think about the questions individually.

A second limitation concerns our use of the final course exam as a measurement instrument. We decided to rely on exam performance because we wished to go beyond short-term effects that would be evident immediately in class after working with a ConcepTest (or a small set of such). To this end, the exam, beyond its function as an indicator of academic success, was specifically designed to require substantial conceptual knowledge. Evaluating learning through exams may, however, be seen as fundamentally problematic, even when the exam content is aligned with the intended conceptual learning outcome: The traditional assessment setup of a final exam involves a number of factors, such as time pressure and possible exam anxiety, that may influence how students can evidence their learning. These factors are not (or much less) present in the active-learning environment in which ConcepTests were used. Hence, general reservations about final exams as measures of learning can certainly be applied to our study as well, and it would therefore be desirable to have alternative measurement methods available that mitigate these issues.

## Conclusion

Our study investigated two different uses of ConcepTests in tutorial group meetings in a course on Real Analysis: The groups worked with the same ConcepTests, with peer discussion being a central element in one group and a more teacher-centered approach in the other group. So the difference between the two groups does not lie in the tasks (and thus, not in the task design), but entirely in the instructional implementation of the ConcepTests. As discussed, our findings show that no significant difference in learning outcome – as measured by the results in two final exams – resulted from the difference between the two groups. Also, both groups equally appreciated the learning opportunities provided for them. Thus, as a central finding of this study, we conclude that, contrary to common belief, it should not be taken for granted that peer discussion is the decisive factor for the positive effects of PI.

The fact that no difference in learning outcome was detected does, of course, not imply that no difference whatsoever exists between the two scenarios. First, it is well conceivable that specific subgroups of students might benefit more from one variant than from the other or that they might appreciate one over the other. It is an intriguing challenge for further research to identify potential characteristics of students that might be relevant in this respect. Secondly, one might hypothesize that the fact that group A practiced peer discussion may result in a difference in certain aspects of argumentation and communication competencies. Such differences would have to be different from those that students were able to evidence in the proof problems that had been posed in the exam. It would be exciting to gain more in-depth understanding of what, if any, differences in the outcomes between the two groups exist in this respect.

**Acknowledgements** The research originated in the WiGeMath project funded by the German Federal Ministry for Education and Research (BMBF), Grant No 01PB14015B. We are grateful to our colleagues from the project for discussions on the initial design of our research.

**Authors' Contributions** All authors contributed to the writing of this paper equally.

**Funding** Open Access funding enabled and organized by Projekt DEAL. German Federal Ministry for Education and Research (BMBF), Grant No 01PB14015B.

## Declarations

**Conflicts of Interest** None.

## References

Balta, N., Michinov, N., Balyimez, S., & Ayaz, M. F. (2017). A meta-analysis of the effect of Peer Instruction on learning gain: Identification of informational and cultural moderators. *International Journal of Educational Research, 86*, 66–77. https://doi.org/10.1016/j.ijer.2017.08.009

Bauer, Th. (2019a). Peer Instruction als Instrument zur Aktivierung von Studierenden in mathematischen Übungsgruppen. *Math. Semesterberichte, 66*(2), 219–241. https://doi.org/10.1007/s00591-018-0225-8

Bauer, Th. (2019b). *Verständnisaufgaben zur Analysis 1 und 2 – für Lerngruppen, Selbststudium und Peer Instruction*. Springer Spektrum. https://doi.org/10.1007/978-3-662-59703-3

Bauer, T., Biehler, R., & Lankeit, E. (2022) Mini-Aufgaben in mathematischen Übungsgruppen zur Analysis: Charakteristika von Aufgaben und Abstimmungsverhalten von Studierenden. In: Hochmuth, R., Biehler, R., Liebendörfer, M., & Schaper, N. (Eds.). Unterstützungsmaßnahmen in mathematikbezogenen Studiengängen – Eine anwendungsorientierte Darstellung verschiedener Konzepte, Praxisbeispiele und Untersuchungsergebnisse (Working Title). Springer Spektrum.

Biehler, R., Hänze, M., Hochmuth, R., Becher, S., Fischer, E., Püschl, J., & Schreiber, S. (2018). *Lehrinnovation in der Studieneingangsphase „Mathematik im Lehramtsstudium" – Hochschuldidaktische Grundlagen, Implementierung und Evaluation - Gesamtabschlussbericht des BMBF-Projekts LIMA 2013 – Reprint mit Anhängen*. Khdm-Report 18–07. Kassel: Universitätsbibliothek Kassel. Retrieved 19 July, 2021, from https://nbn-resolving.de/urn:nbn:de:hebis:34-2018092556466. https://doi.org/10.17170/kobra-2018111412

Butchart, S., Handfield, T., & Restall, G. (2009). Using Peer Instruction to Teach Philosophy, Logic, and Critical Thinking. *Teaching Philosophy, 32*(1), 1–40. https://doi.org/10.5840/teachphil20093212

Catrambone, R. (1996). Generalizing solution procedures learned from examples. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*(4), 1020. https://doi.org/10.1037/0278-7393.22.4.1020

Chi, M. T. (2009). Active-constructive-interactive: A conceptual framework for differentiating learning activities. *Topics in Cognitive Science, 1*(1), 73–105. https://doi.org/10.1111/j.1756-8765.2008.01005.x

Cortright, R. N., Collins, H. L., & DiCarlo, S. E. (2005). Peer instruction enhanced meaningful learning: Ability to solve novel problems. *Advances in Physiology Education, 29*(2), 107–111. https://doi.org/10.1152/advan.00060.2004

Cronhjort, M., Filipsson, L., & Weurlander, M. (2018). Improved engagement and learning in flipped-classroom calculus. *Teaching Mathematics and Its Applications: An International Journal of the IMA, 37*(3), 113–121. https://doi.org/10.1093/teamat/hrx007

Crouch, C. H., & Mazur, E. (2001). Peer instruction: Ten years of experience and results. *American Journal of Physics, 69*(9), 970–977. https://doi.org/10.1119/1.1374249

Crouch, C. H., Watkins, J., Fagen, A. P., & Mazur, E. (2007). Peer instruction: Engaging students one-on-one, all at once. *Research-Based Reform of University Physics, 1*(1), 40–95.

DeLozier, S. J., & Rhodes, M. G. (2017). Flipped classrooms: A review of key ideas and recommendations for practice. *Educational Psychology Review, 29*(1), 141–151. https://doi.org/10.1007/s10648-015-9356-9

Di Martino, P., & Gregorio, F. (2018). The first-time Phenomenon: Successful Students' Mathematical Crisis in secondary-tertiary Transition. In Bergqvist, Österholm, Granberg, & Sumpter (Eds.), *Proceedings of the 42nd Conference of the International Group for the Psychology of Mathematics Education* (Vol. 2, pp. 339–346). PME.

Duncan, D. (2005*). Clickers in the classroom: How to enhance science teaching using student response systems.* Pearson.

Fagen, A. P., Crouch, C. H., & Mazur, E. (2002). Peer instruction: Results from a range of classrooms. *The Physics Teacher, 40*(4), 206–209. https://doi.org/10.1119/1.1474140

Fonseca, B. A. & Chi, M. T. (2011). Instruction based on self-explanation. In: R.E. Mayer, P.A. Alexander (Eds.), *Handbook of research on learning and instruction* (pp. 296–321). https://doi.org/10.4324/9780203839089.ch15

Geisler, S. (2020). *Bleiben oder Gehen? Eine empirische Untersuchung von Bedingungsfaktoren und Motiven für frühen Studienabbruch und Fachwechsel in Mathematik.* (Dissertation), Ruhr-Universität Bochum, Bochum. Retrieved 19 July, 2021, from https://hss-opus.ub.ruhr-uni-bochum.de/opus4/frontdoor/index/index/docId/7163. https://doi.org/10.13154/294-7163

Gok, T. (2012). The effects of peer instruction on students' conceptual learning and motivation. *Asia-Pacific Forum on Science Learning and Teaching, 13*(1).

Gok, T., & Gok, O. (2016). Peer instruction in chemistry education: Assessment of students' learning strategies, conceptual learning and problem solving. *Asia-Pacific Forum on Science Learning and Teaching, 17*(1).

Götz, T. (2004). *Emotionales Erleben und selbstreguliertes Lernen bei Schülern im Fach Mathematik*. Utz.

Hochmuth, R., Biehler, R., Schaper, N., Kuklinski, C., Lankeit, E., Leis, E., Liebendörfer, M., Schürmann, M. (2018). Wirkung und Gelingensbedingungen von Unterstützungsmaßnahmen für mathematikbezogenes Lernen in der Studieneingangsphase: Schlussbericht: Teilprojekt A der Leibniz Universität Hannover, Teilprojekte B und C der Universität Paderborn: Berichtszeitraum: 01.03.2015–31.08.2018. TIB. https://doi.org/10.2314/KXP:1689534117

Kay, R. H., & LeSage, A. (2009). Examining the benefits and challenges of using audience response systems: A review of the literature. *Computers & Education, 53*(3), 819–827. https://doi.org/10.1016/j.compedu.2009.05.001

Kuklinski, C., Leis, E., Liebendörfer, M., Hochmuth, R., Biehler, R., Lankeit, E., Neuhaus, S., Schaper, N., & Schürmann, M. (2018). Evaluating Innovative Measures in University Mathematics – The Case of Affective Outcomes in a Lecture focused on Problem-Solving. In V. Durand-Guerrier, R. Hochmuth, S. Goodchild & N. M. Hogstad (Eds.), *PROCEEDINGS of INDRUM 2018 Second conference of the International Network for Didactic Research in University Mathematics* (pp. 527–536). University of Agder and INDRUM. https://doi.org/10.1007/s40753-019-00103-7

Liebendörfer, M., Göller, R., Biehler, R., Hochmuth, R., Kortemeyer, J., Ostsieker, L., Rode, J., & Schaper, N. (2020). LimSt – Ein Fragebogen zur Erhebung von Lernstrategien im mathematikhaltigen Studium. *Journal Für Mathematik-Didaktik, 42*(1), 25–59. https://doi.org/10.1007/s13138-020-00167-y

Liebendörfer, M., Göller, R., Gildehaus, L., Kortemeyer, J., Biehler, R., Hochmuth, R., Ostsieker, L., Rode, J., & Schaper, N. (2022). On the Role of Learning Strategies for Performance in Mathematics Courses for Engineers. *International Journal of Mathematical Education in Science and Technology.* Online first: https://doi.org/10.1080/0020739X.2021.2023772

Lucas, A. (2009). Using peer instruction and i-clickers to enhance student participation in calculus. *Primus, 19*(3), 219–231. https://doi.org/10.1080/10511970701643970

Marsh, H. W. (1987). The Big-Fish-Little-Pond Effect on Academic Self-Concept. *Journal of Educational Psychology, 79*(3), 280–295. https://doi.org/10.1037/0022-0663.79.3.280

Mazur, E. (1996). *Peer Instruction: A User's Manual*. Prentice Hall.

Mazur, E. (1997). Peer Instruction: Getting students to think in class. In: E.F. Redish, J.S. Rigden (Eds.), *The Changing Role of Physics Departments in Modern Universities* (pp. 981–988). The American Institute of Physics. https://doi.org/10.1063/1.53199

Miller, R. L., Santana-Vega, E., & Terrell, M. S. (2006). Can good questions and peer discussion improve calculus instruction? *Problems, Resources, and Issues in Mathematics Undergraduate Studies, 16*(3), 193–203. https://doi.org/10.1080/10511970608984146

Nicol, D. J., & Boyle, J. T. (2003). Peer instruction versus class-wide discussion in large classes: A comparison of two interaction methods in the wired classroom. *Studies in Higher Education, 28*(4), 457–473. https://doi.org/10.1080/0307507032000122297

Pilzer, S. (2001). Peer instruction in physics and mathematics. *Problems, Resources, and Issues in Mathematics Undergraduate Studies, 11*(2), 185–192. https://doi.org/10.1080/10511970108965987

Porter, L., Bailey Lee, C., Simon, B., & Zingaro, D. (2011). Peer instruction: do students really learn from peer discussion in computing? *In Proceedings of the seventh international workshop on Computing education research* (pp. 45–52). ACM. https://doi.org/10.1145/2016911.2016923

Rach, S. (2014). *Charakteristika von Lehr-Lern-Prozessen im Mathematikstudium: Bedingungsfaktoren für den Studienerfolg im ersten Semester*. Waxmann.

Ramm, G., Prenzel, M., Baumert, J., Blum, W., Lehmann, R., Leutner, D., Neubrand, M., Pekrun, R., Rolff, H. -G., & Rost, J. (2006). *PISA 2003: Dokumentation der Erhebungsinstrumente*. Waxmann.

Rao, S. P., & DiCarlo, S. E. (2000). Peer instruction improves performance on quizzes. *Advances in Physiology Education, 24*(1), 51–55. https://doi.org/10.1152/advances.2000.24.1.51

Renkl, A., & Atkinson, R. K. (2007). Interactive learning environments: Contemporary issues and trends. An introduction to the special issue. *Educational Psychology Review, 19*, 235–238. https://doi.org/10.1007/s10648-007-9052-5

Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin, 140*(6), 1432. https://doi.org/10.1037/a0037559

Schiefele, U., Krapp, A., Wild, K. -P., & Winteler, A. (1993). Der Fragebogen zum Studieninteresse (FSI). *Diagnostica, 39*(4), 335–351.Schöne, C., Dickhäuser, O., Spinath, B., & Stiensmeier-Pelster, J. (2002). *Skalen zur Erfassung des schulischen Selbstkonzepts SESSKO*. Hogrefe.

Schöne, C., Dickhäuser, O., Spinath, B., & Stiensmeier-Pelster, J. (2002). Skalen zur Erfassung des schulischen Selbstkonzepts SESSKO. Bern; Göttingen: Hogrefe.

Simelane, S., & Skhosana, P. M. (2012). Impact of clicker technology in a mathematics course. *Knowledge Management & E-Learning: An International Journal, 4*(3), 279–292. https://doi.org/10.34105/j.kmel.2012.04.023

Smith, M. K., Wood, W. B., Adams, W. K., Wieman, C., Knight, J. K., Guild, N., & Su, T. T. (2009). Why peer discussion improves student performance on in-class concept questions. *Science, 323*(5910), 122–124. https://doi.org/10.1126/science.1165919

Tall, D., & Vinner, S. (1981). Concept image and concept definition in mathematics with particular reference to limits and continuity. *Educational Studies in Mathematics, 12*(2), 151–169. https://doi.org/10.1007/bf00305619

Terrell, M. (2003). Asking good questions in the mathematics classroom. *Mathematicians and Education Reform Forum Newsletter* 15(2) 3–5.

Turpen, C., & Finkelstein, N. (2009). Not all interactive engagement is the same: Variations in physics professors' implementation of peer instruction. *Physical Review Special Topics, Physics Education Research, 5*(2), 1–19. https://doi.org/10.1103/physrevstper.5.020101

Vickrey, T., Rosploch, K., Rahmanian, R., Pilarz, M., & Stains, M. (2015). Research-based implementation of peer instruction: a literature review. *CBE Life Sci Educ, 14*(1), es3. https://doi.org/10.1187/cbe.14-11-0198

Wentzel, K. R. & Watkins, D. E. (2011). Instruction based on peer interactions. In: R.E. Mayer, P.A. Alexander (Eds.), *Handbook of research on learning and instruction* (pp. 322–343). Routledge. https://doi.org/10.4324/9780203839089.ch16

Wolf, K., Nissler, A., Eich-Soellner, E., & Fischer, R. (2014). Mitmachen erwünscht-aktivierende Lehre mit Peer Instruction und Just-in-Time Teaching. *Zeitschrift Für Hochschulentwicklung, 9*(4), 131–153. https://doi.org/10.3217/zfhe-9-04/09

Zingaro, D. (2014). *Peer instruction contributes to self-efficacy in CS1.* Proceedings of the 45th ACM technical symposium on Computer science education - SIGCSE '14, New York: ACM Press, 373–378. https://doi.org/10.1145/2538862.2538878

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## Authors and Affiliations

**Thomas Bauer**[1] · **Rolf Biehler**[2] · **Elisa Lankeit**[2]

Rolf Biehler
biehler@math.uni-paderborn.de

Elisa Lankeit
elankeit@math.uni-paderborn.de

[1] FB Mathematik Und Informatik, Philipps-Universität Marburg, Marburg, Germany

[2] Institut Für Mathematik, Universität Paderborn, Paderborn, Germany