Research article

Comparative judgement as a learning tool in university mathematics: Students' views of benefits and drawbacks

N. Larson[1*]

[1] University of Agder, *Corresponding author. E-mail: niclas.larson@uia.no

Abstract: Comparative judgement (CJ) draws on the idea that it is easier to judge which of two objects weighs more, than to judge the weight of one object. In education, CJ can be used to rank students' responses to a task, as an alternative to assessing each response due to marking rubrics. Research has shown rankings made by students to be valid and reliable, and hence possible to use as a base for summative assessment. Moreover, research has indicated that assessment by CJ can serve as a learning activity for students, i.e., enhance students' learning. This paper reports from an exercise in a calculus course, where the students judged each other's responses by CJ. One of the purposes of the exercise was to explore whether the CJ-process would provide a learning opportunity for the students. The exercise was compulsory, but the results did not count towards their final grade. First, the students were required to respond to the conceptual task "How would you describe the derivative?" Their one-page responses were uploaded to the web engine *No More Marking* (NMM). NMM randomly selects pairings of responses, where students by a mouse click should judge which response shows the best understanding of the derivative. Each student should fulfil at least 11 such pairwise judgements. The research data contain students' responses to the task ($N = 64$), output from NMM based on students' judgements ($N = 61$), and student interviews ($N = 5$). However, the results of this paper are based mainly on data generated by the five interviewees. Excerpts from the interviews support that CJ can improve students' learning. Despite this, output from NMM shows the interviewees spent rather short time on their judgements. This ambiguity reveals the question of whether active participation in CJ can improve students' learning still needs further exploration.

Keywords: comparative judgement, university mathematics, peer assessment, students' engagement

1 Introduction

It is common that mathematics education at university level consists of lecturing in large groups, where the students are passive receivers. However, research has shown that a teaching practice that enables the students to participate more actively in their education may be beneficial for their learning (Freeman et al., 2014). This study originates from when the author followed the course "University level mathematics teaching course", arranged by the Centre for Research, Innovation and Coordination of Mathematics Teaching (MatRIC). One assessment task was to create and try out a teaching/learning activity in a university mathematics course. In line with the outcomes presented by Freeman et al. (2014), the author aimed to do something that would support students' active learning. Since one of the course leaders had introduced the author to *comparative judgement* (CJ), he decided to design an exercise where the students in "Calculus 1" should use CJ. In CJ, scripts (responses) are evaluated by repeated pairwise comparisons, where the judge decides which of two scripts provides the best response. This yields a ranking of the scripts. In this exercise (see the methods section), the students first answered a test task, and then applied CJ to evaluate the scripts handed in.

Previous research has indicated that reflecting on and evaluating peers' responses might be beneficial for students' learning (Jones & Alcock, 2014). Thus, the purpose of this study was to explore students' reflections on their learning through the whole exercise, including both the test and the subsequent CJ-process. Outcomes from the judging process will also be discussed. This paper, which is an elaboration of a paper presented at the MNT Conference 2019 (Larson, 2019), draws mainly on data from follow-up interviews with five students and their judgements of the Calculus class's responses to the test task. The study was steered by the research questions:

> What learning opportunities do students perceive in an exercise where a test is followed by comparative judgement of the replies to the test task?
>
> How do students perform in an exercise where they judge their peers' responses to a test task by comparative judgement?

In the second research question, a student's performance comprehends the quality of their CJ and how much effort they put into their CJ-process.

The aim of the paper is to provide some insights into how CJ can be used for learning, rather than to draw general conclusions. Nevertheless, the outcomes may be valuable, e.g., for teachers in mathematics and mathematics education.

2 Comparative judgement

The idea behind comparative judgement (CJ) draws on the *law of comparative judgement* (Thurstone, 1927), which can be illustrated by an example (cf. Jones & Sirl, 2017, p. 148). If one has a stone of about the size of a fist, it is not very easy to estimate the weight of that stone, with let us say a maximum error of 50 grams. However, if one instead should judge which one of two stones weighs most, the task probably gets easier, even in cases where the difference is less than 50 grams. Transferred to assessment in education, this implies it often is easier to judge which of two responses is the best, than to evaluate the

qualities of one single response (e.g. Goossens & De Maeyer, 2018; Jones & Alcock, 2014).

Adequate and specified assessment criteria are often emphasised as important for a proper assessment (for a brief overview, see Jones & Alcock, 2014). This is the function of marking rubrics, when a solution is marked for example by awarding points. However, since CJ draws on comparisons between two solutions, criteria are superfluous, even though criteria can be used as support also in CJ. One reason for using CJ in mathematics education is that it can be difficult to assess 'conceptual tasks' or 'open-ended questions' in an absolute way. Standard tasks like 'Find the local maximums of $f(x) = x^3 - 5x^2 + 7$ on the interval [−3, 6]', can often successfully be assessed with the support of marking rubrics. However, to assess the task 'Explain how to find the local maximums of a function f, which is defined and continuous on a finite, closed interval [a, b]', rubrics might be less appropriate. The latter task rather suggests an assessment by "direct, holistic and subjective comparisons of the quality of students' work" (Jones & Sirl, 2017, p. 148). For that task, a ranking of the students' responses can be used as support in the assessment process. An example of how to use CJ for assessment will follow here.

Jones and Sirl (2017) assigned a task about a piecewise defined real function of two variables, where the students should describe the properties of the function. They could use e.g. words, symbols, diagram, or a combination of them, to evaluate limits, continuity, partial derivatives, etc. The students should give their response in a square drawn on an A4-page. The scripts were anonymised, scanned, and uploaded to the web engine *No More Marking* (NMM), see www.nomoremarking.com (address valid in October 2025). In NMM, each student should compare students' scripts pairwise and decide which of the two scripts that showed the best conceptual understanding. Each student had to fulfil at least 19 such judgements. Their only decision was to choose which script of a randomly presented pair was the best. No justification was required. This judging ended up in a ranking of the students' scripts, using an algorithm embedded within NMM.

Since the lowest ranked scripts still can be of high quality and vice versa for the highest ranked, this ranking cannot directly replace summative assessment against a *criteria-based* grading scale. Based on the normative ranking from NMM, the examiner needs to decide where to draw the boundaries between the different grades, etc. (see Jones & Alcock, 2014; Jones & Sirl, 2017). That part of the assessment will, however, not be focused on in this paper.

The ranking of the scripts is based on a score calculated by NMM. This score builds on the probability that a script will win a comparison to another script, where this probability stems from the students' judgements (Pollitt, 2012). The reliability of the complete judgement is in NMM given by Scale Separation Reliability (SSR) (Verhavert et al., 2018), which is analogous to Cronbach's alpha (Pollitt, 2012). Cronbach's alpha coefficient >0.7 indicates the result is reliable, while coefficients >0.8 or >0.9 indicate it is highly or very highly reliable (Cohen et al., 2007, p. 506). Earlier, researchers claimed SSR > 0.7 is acceptably high for research on CJ (Bramley & Vitello, 2019). However, more recent research recommends "that researchers should aim for SSR of .8 or greater" (Kinnear et al., 2025, p. 13). Further details about calculation of score and SSR are omitted here.

NMM provides data showing how many judgements each judge (student) fulfilled, the total time spent on the process, the median time for the judgements and the proportion of 'left clicks'. These data give information of the effort each judge has put into the

judgement. There is also a number called 'infit', which measures how consistent a judge's judgements are with all judgements. Lower consistency yields higher infit, and if the infit is too high the judge will be categorised as a 'misfit'. Due to NMM an infit greater than 1.0 indicates the judge has 'some inconsistency', and an infit greater than 1.3 means the judge is 'inconsistent' and thus a misfit. However, an alternative way to identify misfits is that an infit higher than the mean value plus two standard deviations is a misfit (Jones & Sirl, 2017; Pollitt, 2012). Correspondingly, the infit for each script can be calculated. A misfit then means the judges showed low agreement on that script, i.e. it was difficult to judge that script.

3 Previous research

There are several examples of research that promote formative assessment (Black & Wiliam, 2003, 2018) and peer assessment (Gielen et al., 2011; Kollar & Fischer, 2010; Potter et al., 2017; van Zundert et al., 2010) as beneficial for students' learning. Formative assessment given by the teacher aims to support the learning of the student whose work was assessed. However, if formative assessment is given by a peer, that will provide a learning situation for both the assessee and the assessor. Gielen et al. (2011) present five goals of using peer assessment (PA): as learning tool, as a tool for making students participate actively, as assessment tool, as a tool for learning how to assess, and as a social control tool. Although all five goals are of interest, this paper will focus on the first two goals. The main reason for implementing the exercise in the calculus course mentioned earlier was to promote students learning. However, to use PA as a learning tool, will also automatically increase students' active participation in their education.

An important issue about PA is if students are qualified to give proper feedback. That is, do they have enough knowledge in mathematics? Moreover, if they do have enough content knowledge, are they then able to identify the strengths and weaknesses of a solution and give feedback in a way that supports the assessee's learning? It is important that the quality of the feedback is satisfactory, so that the students perceive that PA can be useful and really use their peers' comments (Cho et al., 2006; Gielen et al., 2011; Jones & Alcock, 2014; van Zundert et al., 2010). If the students doubt in the quality of the group's PA, they might be less motivated to put enough effort into their own assessments, or even worse, students who dislike the exercise may deliberately give incorrect feedback (Jones & Alcock, 2014). In addition, mathematics students tend to find comparing peers' answers and giving feedback to peers as less beneficial for their learning than students in English and physics tend to (Potter et al., 2017). Since students' attitudes are likely to affect the qualities and outcomes of PA, this is an important issue. However, although students initially tend to be negative to PA, training and tutoring in PA have a positive effect both on their qualities as assessors and on their opinion (Potter et al., 2017; Van Steendam et al., 2010; van Zundert et al., 2010). This implies exercises including PA should be recurrent. In addition, research has shown that students with insufficient content knowledge anyway might be able to do proper PA. Jones and Sirl (2017) compared the scores of the students' scripts generated by comparative judgement (CJ) with the figures that showed the 'grade of misfit' in their judgements. A high negative correlation would indicate that students performing worse on the test also

would make worse judgements. Even though the correlation, as expected, was negative, it was neither strong ($r = -0.16$) nor statistically significant.

Repeated assessment tends to improve the quality of the judgement. Cho et al. (2006) found assessment due to rubrics made by four peers to be of middling reliability, while assessment made by six peers resulted in excellent reliability. One strength of CJ is that every script is judged several times. In addition, each judgement takes less time with CJ compared to assessment by rubrics. Goossens and De Maeyer (2018) showed the reliability of the CJ to be almost three times as high as for the judgement supported by rubrics, despite the assessment by CJ took less than a third of the time needed for the assessment by rubrics. Furthermore, research has shown the reliability of CJ to be high by measures of the inter-reliability between the judgements made by two expert groups or by two groups of peers (Jones & Alcock, 2014), and since the number of judges counted as 'misfits' was low (Jones & Sirl, 2017). Regarding the validity of CJ, it has been shown to be fair by comparisons between the ranking made by students and a ranking made by experts (Jones & Alcock, 2014; Jones & Sirl, 2017), by comparisons of the outcomes of CJ either with the outcomes from other summative tests in the course (Jones & Alcock, 2014; Jones & Sirl, 2017), or with the students' general level of achievement in mathematics (Bisson et al., 2016). To summarise, there are several indications that CJ generates outcomes of good quality.

CJ has been used in various topics, such as graphic design (Bartholomew et al., 2019), political science (Settembri et al., 2018), to evaluate a written review of a song (Goossens & De Maeyer, 2018), business management (Bouwer et al., 2018), and English, physics and mathematics (Potter et al., 2017). These, and other studies, have also had different purposes for using CJ, like to support learning (Potter et al., 2017), to estimate to what extent a task features problem solving (Holmes et al., 2017), or to evaluate the quality or efficiency of an assessment made by CJ (Bisson et al., 2016; Goossens & De Maeyer, 2018; Jones & Alcock, 2014; Jones & Inglis, 2015; Jones & Sirl, 2017; Tarricone & Newhouse, 2016). Jones and Sirl (2017) also investigated what influenced the students when judging one response to a mathematics task as 'the better'. Their survey showed "Accuracy of answers", "Appropriateness of examples" and "Use of facts and theorems" to be important, while "Flair and originality" and "Quantity of ink used" were seen as less important (p. 155).

Further, CJ has been used to evaluate "the relative benefits of abstract and contextualised representations for introducing key concepts to students" (Jones et al., 2016, p. 2), or "to offer students an opportunity to reflect on their own conceptual understanding and communication of mathematical ideas, and thus to promote higher-order learning" (Jones & Alcock, 2014, p. 1776). An exercise where the students first should answer a 'test question', and then judge peers' scripts by CJ, will provide several learning opportunities. The preparation for the test and the test situation itself offers learning opportunities, as well as when students afterwards scrutiny their peers' scripts. If the students, in addition, are requested to justify their CJ-decisions (e.g. Bartholomew et al., 2019; Goossens & De Maeyer, 2018; Potter et al., 2017), they might also learn "from providing peer feedback (in the 'student as tutor' model) through comparisons and associated comments to explain their reasoning" and "from feedback received from peers on their own work" (Potter et al., 2017, p. 91). On the other hand, if no justifications or comments are required (e.g. Bisson et al., 2016; Jones & Alcock, 2014; Jones & Sirl, 2017), the judgements will take less time and hence the assessor will be able to

scrutinise more scripts and thus get more input from peers' solutions. Hence, there are benefits with both including and not including justifications in CJ.

4 Method

The students participating in the study followed Calculus 1, a 15 ECTS points course at university mathematics beginners' level. The students followed different study programmes, e.g. there were students aiming for a master's degree in mathematics or mathematics education, students planning to study mathematics for one year, and secondary school pre-service teachers. In addition, the students had different experiences of being in higher education. Approximately 70 students followed the course. About half of the students were in their first semester, but some had already finished their master's qualification. The teaching consisted of three lectures and one group session every week; presence was not obligatory. The course was examined by a written exam at the end of the semester. In addition, there were two compulsory exercises during the semester.

The CJ-exercise this paper mainly draws on was one of these compulsory exercises. This exercise consisted of a test and a subsequent assessment by CJ. Neither affected the students' course grade, but they had to fulfil the exercise to be qualified to participate in the final written exam. One week before the test, the test task was presented to the students (cf. Jones & Sirl, 2017). The task was "How would you describe the derivative? You may use e.g. words, graphs, examples, calculations or pictures in your explanation." The students could prepare for the test in any way they wanted to, but during the 20 minutes test they would not be allowed to use anything but pen and paper (and ruler, compass, etc.).

There were 67 students participating in the test. Their responses were scanned and uploaded to No More Marking (NMM). Then, they had eleven days to fulfil at least 11 judgements, with an upper limit of 120 judgements. However, three students did not give permission to participate in research, so their scripts and judgements were deleted from NMM before the analysis. In addition, four students did not fulfil the judging process. Thus, the data from the test include 64 scripts and 61 judges (since one student registered twice).

After the exercise, the students were invited to participate in a follow-up interview. Five students volunteered, which means a convenience sample was employed (Cohen et al., 2007). All five interviewees were female pre-service teachers, and they were interviewed in groups of two and three. The pair, called Interviewee 1 and Interviewee 2, was interviewed in English. The triplet was interviewed in Norwegian, and will henceforth be called Interviewee 3, 4 and 5. The interviews were conducted in a group room at the university. They were semi-structured and followed an interview guide, with the following pre-determined themes:

- The test task
  - Was the task interesting?
  - How did you prepare for the test?
  - What did you learn from the task?
- The assessment by CJ
  - What aspects affected your judging?
  - What did you learn from other students' scripts?

- Challenges and issues with CJ?

The duration was approximately 15 minutes for the pair and a bit longer for the triplet. The interviews were video recorded and thereafter transcribed by the author. The transcripts were analysed utilising a thematic content analysis (Cohen et al., 2007), supported by the themes from the interview guide. The outcomes of this analysis are presented in subsection 0, including utterances that support the claims. Together with the quantitative data generated by NMM, this means the paper draws on mixed methods.

5 Results

This section begins with a presentation of results from the whole group's assessment in NMM. Although this paper is focused on data from the five interviewees, this provides a backdrop for the results presented in subsections 0 and 0.

In total, the 61 judges made 1433 judgements. A rule of thumb is that the number should be at least 10 times the number of scripts (Bisson et al., 2016), which in this case gave a minimum of 640 judgements. Hence, the number of judgements is enough to give a reliable result. The SSR generated by NMM was 0.88, which implies the output can be used for research (SSR > 0.8) (Kinnear et al., 2025). The mean of the judges' infit was 0.97 and the standard deviation was 0.35, which means an infit greater than 1.67 would indicate a misfit due to the rule 'mean + 2 std.dev.' (Jones & Sirl, 2017; Pollitt, 2012). The outcomes generated by the whole student group will, however, not be discussed in detail in this paper. Here, the main focus will be on the five interviewees' experiences and learning in connection with this exercise and their performance on the CJ-exercise.

5.1 The interviewees' performances on the exercise

Here, an analysis of the interviewees' performance in the CJ-process is presented. The subsection, however, begins with a presentation of two scripts and an analysis of all interviewees' scripts. Even though this is not directly connected to the research questions, it serves as a backdrop for the subsequent results on CJ. As earlier mentioned, the test task was an open-ended question, where the students were asked to describe the derivative. Figures 1 and 2 illustrate the test task and might provide some transparency regarding what aspects that were rewarded by the judges.

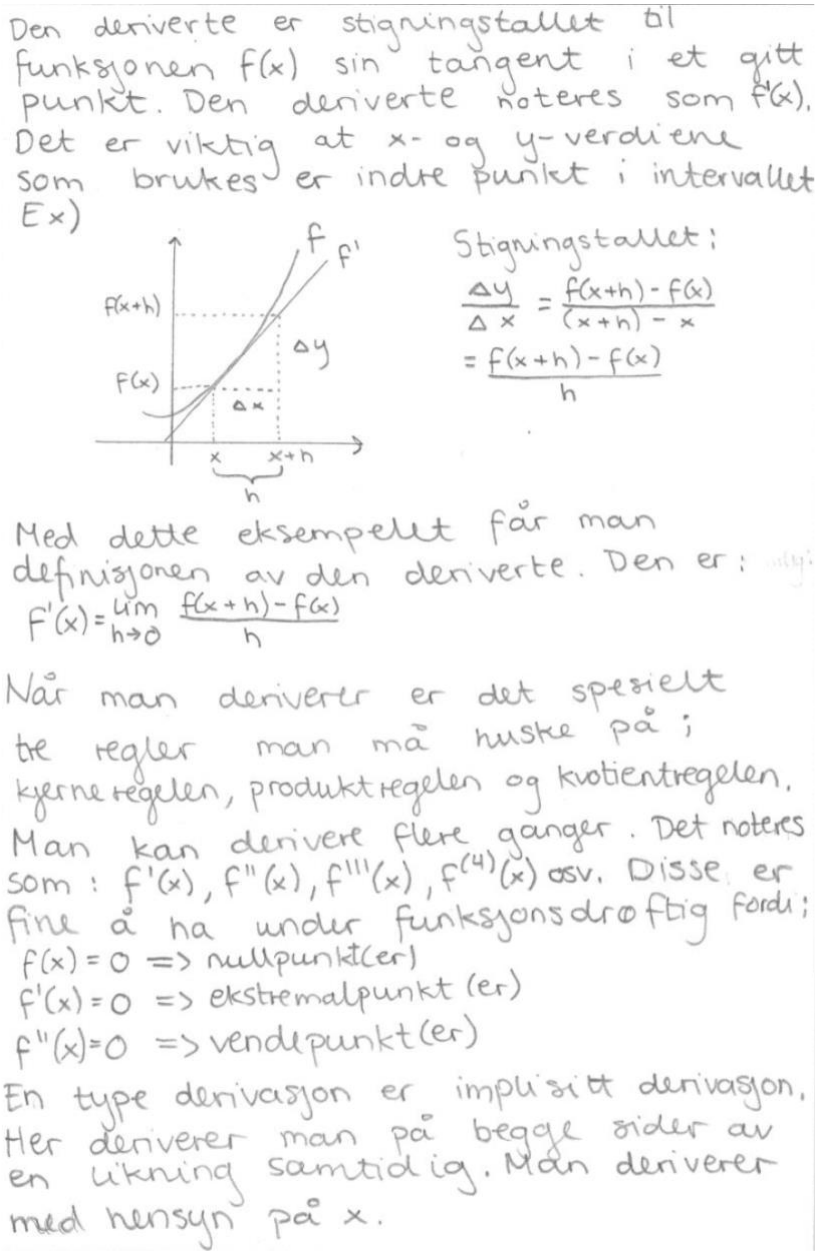

Figure 1. *Interviewee 4's script – ranked 8 of 64 by the whole student group*.

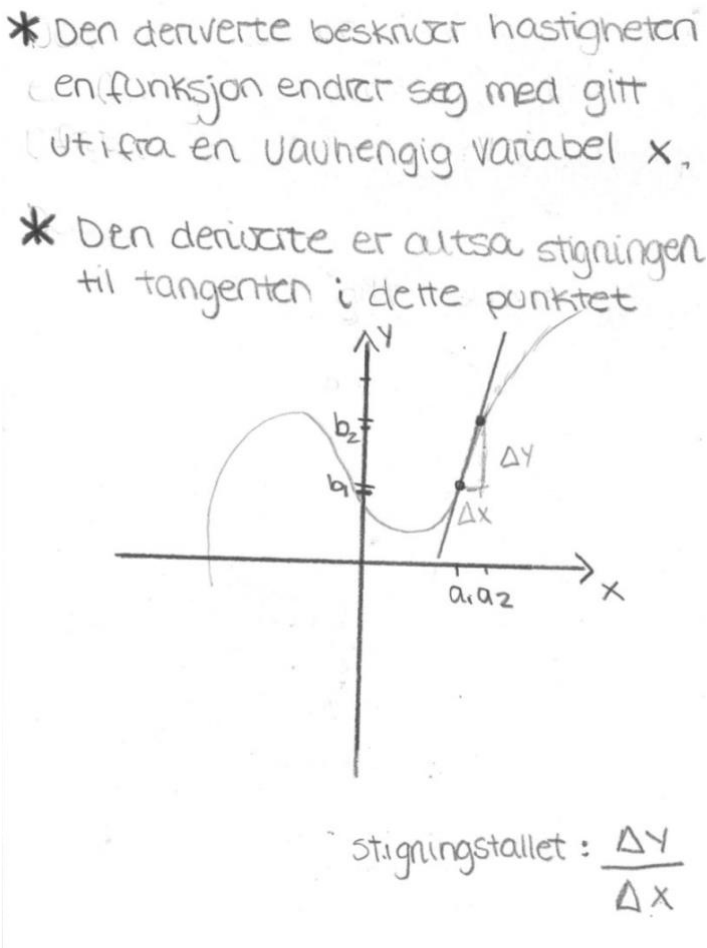

Figure 2. *Interviewee 2's script – ranked 50 of 64 by the whole student group*.

The interviewees' scripts were scrutinised by a content analysis, where the categories were derived directly from the scripts following a grounded theory approach. Table 1 shows the result from this analysis.

Table 1. *Features emerging in the interviewees' scripts*.

Features emerging in the scripts	**I1**	**I2**	**I3**	**I4**	**I5**
Growth rate	X	X	X		X
Slope of e.g. secant, tangent or graph	X	X	X	X	X
The denotation $f'(x)$	X		X	X	X
Definition of derivative through formula	X		X	X	X
Limit, or *h* approaches 0	X		X		
Example deduced through definition, i.e. $(x^2)' = 2x$			X		X
Concretisation, horizontal graph means derivative 0			X		
Differentiation rules, e.g. quotient rule				X	
Higher-order derivatives				X	
Graph(s)	X	X	X	X	X
CJ-ranking by the whole group (out of 64 scripts)	13	50	21	8	17

Table 1 shows the script including the fewest features was lowest ranked of the five scripts. However, the script with most features was not highest ranked. A probable explanation is that it was a bit jumbly and that the writing looked a bit blurry, at least in the scanned copy shown in NMM. The author of this paper made a ranking of these five scrips based on a holistic evaluation before the content analysis revealed which and how many of the features that were present. Then, Interviewee 3's script was ranked as the best. Otherwise, the author's ranking agreed with the ranking made by the whole student group. Table 1 just gives a rough description of the benefits of the scripts. There were, of course, qualitative differences in the presentations as well. A deeper analysis of these is, however, beyond the scope of this paper.

Table 2 presents statistics from the interviewees' judging process, derived directly from NMM. Recall the students were requested to do at least 11 judgements, with an upper limit of 120. Table 2 also shows the median time for the judgements, the proportion of choosing the left script, and the quality measure 'infit' described in chapter 2.

Table 2. *Statistics from the interviewees' judgements*.

	Number of judgements	Median time (s)	Left click	Infit
Interviewee 1	11	34.8	55 %	0.71
Interviewee 2	32	6.6	38 %	1.42
Interviewee 3	46	17.8	50 %	1.23
Interviewee 4	12	89.8	57 %	0.90
Interviewee 5	18	10.0	78 %	1.00

Interviewee 2 had a median time of 6.6 seconds, which indicates she did not spend much time on her judgements. Even if data from NMM (not presented here) show she spent more time on the first judgements, it might not be possible to perform a fair judgement in just a few seconds. In addition, the infit of Interviewee 2 was high (1.42). Remember, judges with an infit greater than 1.3 is a misfit due to the suggestions in NMM, while the rule 'mean + 2 std.dev.' (Jones & Sirl, 2017; Pollitt, 2012) here gives 1.67 as the limit for misfits.

Further, Table 2 shows Interviewee 5 had 78 % left clicks in her 18 judgements. That proportion is rather high, but for 18 judgements it might be by coincidence rather than by negligence, although the median time 10.0 seconds indicates her judgements not being very ambitious. Table 2 also shows that Interviewee 1 fulfilled the minimum number of judgements and Interviewee 4 just one more than the minimum. However, their median time and infit both indicate they put more effort into the judgements they made than the other three interviewees. It is noteworthy that the actual number of judgements can be a bit higher, since some judgements might have disappeared when three scripts were deleted. A curiosity is that, in total, the five interviewees were assigned to judge their own scripts three times. In two of these judgements, the judge favoured her own script.

5.2 Student interviews

The first interview question was how the students found the task they should solve during the test. The interviewees said the task was different compared to 'standard tasks', and they were mainly positive to the task.

Interviewee 2: I think it was a really nice task, because we could describe the derivative in different ways, like you could draw it or you could explain, or you could do both.

Interviewee 4: It covers much of what we have learnt. That was good. You will get some explanations of what's behind, and not just find the derivative to some function.

Interviewee 3: I agree [with Interviewee 4]. It is good to be forced to see what it's all about.

Their way of preparing for the test varied, and they used different resources. For example, they read in the course book, used Google to find pictures, watched YouTube, used their lecture notes, and one interviewee said she practised on how to compress the content since the space for their response was limited to less than one A4-page in the test situation. Mainly, they found it positive to know the task in advance, but that could also have led to less preparation.

Interviewee 2: The answer gets better, but actually, I think we would prepare more if we didn't know it. Because then you have to read on different ...

Regarding the CJ-part, the interviewees were asked what guided them in their judging process in NMM. All interviewees referred to the content of the scripts as important. The layout of the scripts also influenced their judgement, as well as if the handwriting was clear or not.

Interviewee 2: Maybe I'm weird, but I actually look at how people write too. Because I feel like if they write really ugly, then it's like hard to read. But, of course, I think it's nice that people use like different examples like that they paint and also write formulas.

Interviewee 1: I also saw how some people went in depth with the details and so on, while some other just described it very easily. And the one who went in depth showed that they really knew what it was. So, that was kind of my way of judging which one was the best.

Interviewee 5: It was which had the best content, but at the same time, you are a little affected by the words and the structure in a way. If there is a lot of clutter, I feel like, it looks much better with one that is neat, and I feel more like clicking on that.

The interviewees experienced the CJ-process was beneficial for their learning, and that they learnt through reading peers' scripts.

Interviewee 2: This was nice, because when I answered my answer, and then I saw the others' answers, I learned really much.

Interviewee 2: I was like sitting on the screen taking pictures of the other people's answers.

Interviewee 1: Really different ways to explain the derivative. How some people had written a lot, and then some people didn't write so much, but used symbols and mathematics and it was kind of the same thing.

Interviewee 4: Because automatically it will be that when you read the others, you will compare it to what you have written yourself, and to find out should I have included that, should I have used another formulation.

The interviewees were mainly positive to the exercise, but admitted they would have prepared more and made more effort if the result also had affected their grade. That could also have increased their learning because the quality of the scripts they should judge would possibly have been higher. They were positive to include exercises with CJ in other courses, but they also said they missed that they could not give feedback to the scripts they judged, and they mentioned the case where both scripts are poor.

Interviewee 2: I think I like more the exercise with peer assessment [a voluntary exercise during the course], it's like not the same, but then you go like deeper in and can comment on something and not just press which one you like the best.

Interviewee 1: And even when you're pressing like that one answer is better than the other, it doesn't necessarily mean that the one you said was better is right.

To summarise, the interviewees found the exercise interesting and beneficial for their learning, especially the CJ-process where they learnt from peers' scripts by scrutinising them.

6 Discussion

The interviewees saw the test task as different from regular tasks like 'find the derivative of …'. They expressed they were positive to the task and that this helped them understand the derivative in a deeper way. The interviewees also claimed that they learnt 'a lot' by looking at peers' scripts during the judging process. Interviewee 2 even took pictures of other scripts, since she found them valuable for her learning. This indicates that exercises with CJ can be beneficial for students' learning; at least the interviewees experienced they learnt from the exercise (cf. Jones & Alcock, 2014). Nevertheless, although Interviewee 2 found she learnt a lot from the CJ-process, the median time for her 32 judgements was 6.6 seconds, which shows she could not have used every judgement for learning. However, she did spend more than 30 seconds on five judgements, so her claim in the interview might refer to these judgements. To summarise the statistics regarding the judgements (Table 2), three of the interviewees fulfilled less than 20 judgements and the two interviewees who fulfilled more than 30 judgements both had a rather high infit. This indicates the interviewees did not put much effort into their judgements, even though they claimed it was beneficial for their learning. The conclusion that the interviewees did not make enough effort might be incorrect, but still the statistics indicates the interviewees did not make use of CJ as a learning activity to

the extent their positive interview responses suggest they could have. A potential explanation is that the interviewees might have exaggerated their positive impressions "to be unduly helpful by attempting to anticipate what the interviewer wants to hear" or "to show themselves in a good light" (Cohen et al., 2007, p. 153). The matter that the interviewer also was their teacher supports this conclusion, which, at least partly, explains the inconsistent results.

The interviewees claimed that the content was the most important factor when they judged which of two scripts was the best. That is in line with earlier research (Jones & Sirl, 2017), who found that "accuracy of answers" was the most important factor. However, the interviewees also mentioned the layout and handwriting of the script. "Neatness of presentation" seems to be a necessary, but not sufficient, factor for a script to be ranked high, although the mathematical content is more important (cf. Jones & Sirl, 2017). This might be an argument for using marking rubrics rather than holistic comparisons, which may be affected by more aesthetical qualities. However, supportive instructions can be utilised also for CJ, to guide the judges towards a focus on the mathematical aspects. Besides, communication is an important part of mathematics, and a proper layout and clear presentation is a part of the communication. In addition, writing mathematics is one of the basic skills in the syllabus for Norwegian schools (Norwegian Directorate for Education and Training, 2020). Hence, it might be fair also to consider the neatness in the judging process. Though, if CJ is used for summative assessment, technical issues like if the script looks blurry on the screen cannot be allowed to affect the grade. This demonstrates the importance of that the teacher checks and processes the ranking before the grade is decided (Jones & Alcock, 2014; Jones & Sirl, 2017). Furthermore, if CJ is used for summative assessment, it might be necessary to delete the cases where students have judged their own script. In a study where the result also counted towards the grade, the students favoured their own script in all 29 cases they were assigned to judge it (Jones & Alcock, 2014, p. 1780).

The interviewees found it problematic that they in every judgement had to choose one script as winner, even when both scripts were poor. Partly, they also saw that they should not comment on the scripts as defective. Hence, one interviewee said she preferred exercises where they were asked to give constructive feedback to a peer. However, if CJ includes commenting on the solutions, the students will be able to judge fewer scripts, which might decrease their opportunities to learn from scrutinising peers' responses. Since both kinds of exercises have their benefits and drawbacks, this is an argument to in mathematics courses include both CJ-exercises where no comments are expected, and exercises where the students shall give feedback to peers. Even if research has shown mathematics students to be less positive regarding how giving feedback contributes to their own learning (Potter et al., 2017), there are arguments suggesting peer assessment (PA) is beneficial for students' learning (e.g. Gielen et al., 2011). It can also be considered to include CJ-exercises where the students shall justify their decisions (cf. Bartholomew et al., 2019; Goossens & De Maeyer, 2018; Potter et al., 2017). However, some care should be taken not to introduce too many new and different PA-exercises in the same first semester course.

The CJ-exercise focused on in this study was compulsory, but it did not affect the students' final grade. Consequently, the students said, this made them prepare less. A possible conclusion is that students saw the exercise as redundant since it did not affect the grade. To avoid this, the benefits of PA, including CJ, as a *learning opportunity* must

be promoted. As earlier mentioned, frequent training is likely to enhance the quality of the assessment (Potter et al., 2017; Van Steendam et al., 2010; van Zundert et al., 2010), which is important for students' engagement in PA (Cho et al., 2006; Gielen et al., 2011; Jones & Alcock, 2014; van Zundert et al., 2010). One possibility to get the students more engaged is to regularly practice assessment during lectures, to emphasise the benefits of PA and improve the students' assessment skills. Another possibility to increase students' engagement is to include results from the exercises in the final assessment (e.g. Jones & Sirl, 2017), either as a graded course part or to earn bonus points for the final written exam.

7 Concluding remarks

This study supports exercises including CJ can be beneficial for students' learning. However, a clear limitation is that this conclusion is based exclusively on data from five student interviews, where all interviewees were pre-service teachers, selected by a convenience sample. In addition, the interviewer was their course teacher, which increases the risk that the interview outcomes were biased (cf. Cohen et al., 2007, p. 153). The statistics from NMM (Table 2) support, at least partly, that the interviewees were more positive during the interviews than during their CJ-work. Yet, the outcomes from the interviews were quite distinct regarding the interviewees' positive impression regarding CJ as a learning activity and the exercise overall. A possible explanation to the inconsistent results could be that the interviewees actually were very positive to this exercise but exaggerated their effort and overestimated the quality of their CJ. Furthermore, the aim of this paper never was to draw general conclusions, but to provide illustrative examples of how CJ could be used as a learning activity. This reduces the concerns regarding the trustworthiness of the paper.

Regardless of the significance one accredits to the paper's results or its limitations, the potential benefits of CJ as a learning activity needs further investigation (cf. Jones & Sirl, 2017). One suggestion is to first let the whole student group solve two tasks in different topics. Then, split the group into two halves, which assess one task each by CJ. Finally, let the whole group undertake a test with tasks in both topics, to explore if the students will perform better in the topic they worked with in CJ. If results from such a study would show that CJ is beneficial for students' learning, they could be a powerful tool to convince students to engage more actively in their education. This would support mathematics teachers in the continual, important challenge to provide formative learning opportunities and encourage the students to take part in them.

References

Bartholomew, S. R., Zhang, L., Garcia Bravo, E., & Strimel, G. J. (2019). A tool for formative assessment and learning in a graphics design course: Adaptive comparative judgement. *The Design Journal*, *22*(1), 73–95. https://doi.org/10.1080/14606925.2018.1560876

Bisson, M.-J., Gilmore, C., Inglis, M., & Jones, I. (2016). Measuring conceptual understanding using comparative judgement. *Research in Undergraduate Mathematics Education*, *2*(2), 141–164. https://doi.org/10.1007/s40753-016-0024-3

Black, P. & Wiliam, D. (2003). 'In Praise of Educational Research': Formative Assessment. *British Educational Research Journal*, *29*(5), 623–637. https://doi.org/10.1080/0141192032000133721

Black, P. & Wiliam, D. (2018). Classroom assessment and pedagogy. *Assessment in Education: Principles, Policy & Practice*, *25*(6), 551–575. https://doi.org/10.1080/0969594X.2018.1441807

Bouwer, R., Lesterhuis, M., Bonne, P., & De Maeyer, S. (2018). Applying criteria to examples or learning by comparison: Effects on students' evaluative judgment and performance in writing. *Frontiers in Education*, *3*, Article 86, 1–12. https://doi.org/10.3389/feduc.2018.00086

Bramley, T. & Vitello, S. (2019). The effect of adaptivity on the reliability coefficient in adaptive comparative judgement. *Assessment in Education: Principles, Policy & Practice*, *26*(1), 43–58. https://doi.org/10.1080/0969594X.2017.1418734

Cho, K., Schunn, C. D., & Wilson, R. W. (2006). Validity and reliability of scaffolded peer assessment of writing from instructor and student perspectives. *Journal of Educational Psychology*, *98*(4), 891–901. https://doi.org/10.1037/0022-0663.98.4.891

Cohen, L., Manion, L., & Morrison, K. (2007). *Research methods in education.* Routledge.

Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., & Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences*, *111*(23), 8410–8415. https://doi.org/10.1073/pnas.1319030111

Gielen, S., Dochy, F., Onghena, P., Struyven, K., & Smeets, S. (2011). Goals of peer assessment and their associated quality concepts. *Studies in Higher Education*, *36*(6), 719–735. https://doi.org/10.1080/03075071003759037

Goossens, M. & De Maeyer, S. (2018). How to obtain efficient high reliabilities in assessing texts: Rubrics vs comparative judgement. In E. Ras & A. E. Guerrero Roldán (Eds.), *Technology Enhanced Assessment. 20th International Conference, TEA 2017. Barcelona, Spain, October 5–6, 2017. Revised Selected Papers* (pp. 13–25). Springer Nature Switzerland AG. https://doi.org/10.1007/978-3-319-97807-9_2

Holmes, S. D., He, Q., & Meadows, M. (2017). An investigation of construct relevant and irrelevant features of mathematics problem-solving questions using comparative judgement and Kelly's Repertory Grid. *Research in Mathematics Education*, *19*(2), 112–129. https://doi.org/10.1080/14794802.2017.1334576

Jones, I. & Alcock, L. (2014). Peer assessment without assessment criteria. *Educational Research and Evaluation*, *18*, 425–440. https://doi.org/10.1080/03075079.2013.821974

Jones, I. & Inglis, M. (2015). The problem of assessing problem solving: can comparative judgement help? *Educational Studies in Mathematics*, *89*(3), 337–355. https://doi.org/10.1007/s10649-015-9607-1

Jones, I., Inglis, M. Gilmore, C., & Bisson, M.-J. (2016). *Measuring Conceptual Understanding: The Case of Teaching with Abstract and Contextualised Representations.* (Final Report). Nuffield Foundation. Retrieved online 12.08.2025 http://www.nuffieldfoundation.org/sites/default/files/files/MCU_FINALREPORT.pdf

Jones, I. & Sirl, D. (2017). Peer assessment of mathematical understanding. *Nordic Studies in Mathematics Education*, *22*(4), 147–164. https://doi.org/10.7146/nomad.v22i4.148924

Kinnear, G., Jones, I. & Davies, B. (2025). Comparative judgement as a research tool: A meta-analysis of application and reliability. *Behavior Research Methods*, *57*(8), Article 222. https://doi.org/10.3758/s13428-025-02744-w

Kollar, I. & Fischer, F. (2010). Peer assessment as collaborative learning: A cognitive perspective. *Learning and Instruction*, *20*(4), 344–348. https://doi.org/10.1016/j.learninstruc.2009.08.005

Larson, N. (2019). Comparative judgement as a learning activity. *Nordic Journal of STEM Education*, *3*(1), 135–139. [Proceedings for the MNT Conference 2019, UiT – The Arctic University of Norway, Tromsø, Norway, March 28–29, 2019]. https://doi.org/10.5324/njsteme.v3i1.2992

Norwegian Directorate for Education and Training. (2020). *Curriculum for the common core subject of mathematics (MAT1-05)*. Norwegian Directorate for Education and Training. https://www.udir.no/lk20/mat01-05

Pollitt, A. (2012). The method of Adaptive Comparative Judgement. *Assessment in Education: Principles, Policy & Practice*, *19*(3), 281–300. https://doi.org/10.1080/0969594X.2012.665354

Potter, T., Englund, L., Charbonneau, J., Thompson MacLean, M., Newell, J., & Roll, I. (2017). ComPAIR: A new online tool using adaptive comparative judgement to support learning with peer feedback. *Teaching & Learning Inquiry*, *5*(2), 89–113. https://doi.org/10.20343/teachlearninqu.5.2.8

Settembri, P., Van Gasse, R., Coertjens, L., & De Maeyer, S. (2018). Oranges and apples? Using comparative judgement or reliable briefing paper assessment in simulation games. In P. Bursens, V. Donche, D. Gijbels, & P. Spooren (Eds.), *Simulations of decision-making as active learning tools: Design and effects of political science simulations* (pp. 93–108). Springer International Publishing AG. https://doi.org/10.1007/978-3-319-74147-5_8

Tarricone, P. & Newhouse, P. (2016). Using comparative judgement and online technologies in the assessment and measurement of creative performance and capability. *International Journal of Educational Technology in Higher Education*, *13*(16), 1–11. https://doi.org/10.1186/s41239-016-0018-x

Thurstone, L. L. (1927). A law of comparative judgement. *Psychological Review*, *34*, 273–286. http://dx.doi.org/10.1037/h0070288

van Zundert, M., Sluijsmans, D., & van Merriënboer, J. (2010). Effective peer assessment processes: Research findings and future directions. *Learning and Instruction*, *20*(4), 270–279. https://doi.org/10.1016/j.learninstruc.2009.08.004

Van Steendam, E., Rijlaarsdam, G., Sercu, L., & Van den Bergh, H. (2010). The effect of instruction type and dyadic or individual emulation on the quality of higher-order peer feedback in EFL. *Learning and Instruction*, *20*(4), 316–327. https://doi.org/10.1016/j.learninstruc.2009.08.009

Verhavert, S., De Maeyer, S., Donche, V., & Coertjens, L. (2018). Scale separation reliability: What does it mean in the context of comparative judgment? *Applied Psychological Measurement*, *42*(6), 428–445. https://doi.org/10.1177/0146621617748321