Comparative judgement and its impact on the quality of students' written work in mathematics

Jennifer Palisse, Deborah King, Mark Maclean

Comparative judgement and its impact on the quality of students' written work in mathematics

Jennifer Palisse[1], Deborah King[1], and Mark MacLean[2]

[1]The University of Melbourne, Faculty of Science, Australia, jmpalisse@student.unimelb.edu.au; [2]University of British Columbia, Canada

In this experimental study, we investigated the use of comparative judgement as a way to facilitate learning through comparison. Secondary and undergraduate mathematics students (N = 24) evaluated peer work on the topic of solving rational inequalities where solutions were presented to them either in pairs or one-at-a-time. Presenting solutions in pairs resulted in a greater improvement in performance outcomes than presenting solutions individually. Students who compared solutions were more focused on how mathematics was communicated, rather than the final answer, and were more likely to implement 'good' features into their own work. Comparative judgement may therefore be effective for improving performance outcomes because it facilitates the noticing and implementation of 'good' features.

Keywords: Novel approaches to teaching, assessment practices in university mathematics education, comparative judgement, evaluative judgement, learning by evaluating.

INTRODUCTION

In our daily lives, the act of comparison is integral to the decision-making process. This fundamental mechanism extends to the realm of learning, including the context of mathematics education (Alfieri et al., 2013). There are several ways to incorporate comparison within mathematics education, for example, engaging students in discussions where they share problem-solving strategies (Boaler, 1998), using direct instruction that presents side-by-side examples with explicit emphasis on comparisons (Begolli & Richland, 2016), or offering students pairs of worked examples to facilitate self-guided comparison of strategies (Star & Rittle-Johnson, 2009). This study introduces an alternative approach known as comparative judgement, where students assess pairs of peer responses to a question. Responses are presented as pairs and students judge which of the two is 'better'. Students complete multiple rounds of comparisons enabling the responses to be ranked from 'best' to 'worst'.

This study builds on empirical findings which highlight improved performance outcomes for students using comparative judgement. Bartholomew et al. (2019) demonstrated that middle-school students who engaged in comparative peer-based assessment while designing travel brochures outperformed those who participated in a face-to-face peer feedback exercise. In a study with first-year design students generating Point of View statements, Bartholomew et al. (2022) observed that a brief 20-minute comparative judgement task resulted in higher-quality Point of View statements compared with students who did not engage in comparison but were instead given extra time to complete their work. In the context of English, Bouwer et al. (2018)

found that students engaging in comparative judgement produced higher-quality essays compared to those who evaluated peer work sequentially using marking criteria.

While these studies show some promise for improved learning outcomes in a variety of subjects, there is a gap in research addressing whether similar effects are observed in the context of mathematics.

LEARNING BY COMPARING

Since a theoretical framework for learning through comparative judgement has not yet been established, we draw upon the research tradition of learning from worked examples to explain why comparing might be useful for learning. In this paper, the term *worked examples* refers to examples written by educators for the purpose of learning. We use the term *worked solutions* to refer to solutions written by students for the purpose of assessment. Worked solutions may include traits such as scribbling out, hard-to-follow layout, or poorly worded explanations, possibly making them more difficult for students to learn from than a carefully designed worked example.

Learning from worked examples involves providing learners with a problem, the steps that were taken to reach a solution, as well as the final solution. Presenting students with multiple worked examples simultaneously is more effective for learning than providing the same worked examples sequentially (Alfieri et al., 2013). Comparing worked examples facilitates the recognition of underlying structures by enabling learners to notice commonalities across multiple examples which can then be applied to future problems with similar features. According to variation theory, the ability to identify these distinctions is necessary for learning (Marton, 2015). In order to generate new knowledge, one must notice a new aspect which can only occur when it is contrasted against previously noticed aspects in a pattern of variation. The process of comparison becomes a useful means to generate such variations.

While current research argues that comparing worked examples is better for learning than learning from examples one-by-one, it remains unclear whether this holds true in the context of comparative judgement. In comparative judgement, the emphasis is on *evaluating* worked solutions rather than *understanding* worked examples. Additionally, while including variation in worked examples is beneficial, too much variation might exist across student-produced worked solutions. When too much variation is present and both relevant and irrelevant elements vary simultaneously, it can be more difficult for learners to discern relevant information whilst simultaneously ignoring irrelevant information (Marton, 2015).

That said, certain aspects of comparative judgement do align with recommended pedagogical practices associated with learning from worked examples. First, learners are unlikely to notice similarities and differences across multiple solutions without a prompt to do so (Alfieri et al., 2013). Comparative judgement requires students to explicitly compare two solutions, thereby increasing the likelihood that they recognise similarities and differences between solutions. Second, during comparative judgement, students complete a number of comparisons which can generate variety across

solutions. Increased variety should increase the chance of students noticing similarities and differences and exposure to multiple approaches should increase procedural flexibility, that is, the ability to select and apply different procedures effectively (Große, 2014). Lastly, Seery and Canty (2018) argue that comparative judgement can support self-reflection and self-regulation. By comparing multiple pieces of work, students can position their own performance against those of others, providing them with a better understanding of the quality of their own work. This requires students to establish their own criteria for proficiency, especially if marking schemes are not provided, which helps students build an understanding of what it means to be capable.

CURRENT STUDY

The current study extends previous literature by examining learning through comparative judgement in the context of mathematics. While our focus is on comparing the effectiveness of different instructional methods in mathematics learning, our approach aligns with the broader goal of comparative judgement research wherein our interest lies in investigating the impact of comparative judgement on the overall quality of students' work. In the context of mathematics, this will likely relate to mathematical proficiency, ability to communicate ideas, and appropriateness of solution methods.

For the current study, students participated in a peer review activity and were randomly assigned to one of two groups. The first group evaluated other students' solutions to a rational inequality problem presented in pairs (*compare* group) while the second group evaluated the same set of solutions one-at-a-time (*sequential* group). We hypothesised that the compare group would outperform the sequential group as current literature argues that comparing worked examples is better for learning than studying worked examples one-at-a-time. Additionally, prompting students to compare worked examples was shown to be effective for learning, and we hypothesised that asking students to select which of two solutions was 'better' would have the same effect as an explicit prompt to compare.

Despite evidence that comparing examples is effective for learning, it is unclear whether this approach extends to the context of comparative judgement. When learning from worked examples, educators carefully create examples to facilitate understanding, while in comparative judgement, solutions created by other students lack intentional instructional design. Additionally, when comparing examples, educators purposefully select complementary examples that make similarities and differences more noticeable. During comparative judgement, when pairs are selected by a computer, pairings are not selected with purposeful variation in mind. For the current study, solutions were not deliberately paired to highlight discernible differences between pairs. Consequently, any potential benefits from comparisons may be negated by the use of worked solutions not designed for instructional purposes or solution pairings that do not emphasise key elements.

METHOD

Participants

The study included 24 participants of which 15 were Year 10 and 11 students from a select entry secondary school (10 female, 5 male) in Victoria, Australia. Students in Year 10 were accelerated students studying mathematics one year ahead of their peers. All Year 10 and Year 11 students were studying the same Year 11 mathematics subject. The remaining nine participants were undergraduate students (1 female, 8 male) who were studying undergraduate mathematics. Secondary school students were unfamiliar with rational inequalities while undergraduate students would not have been shown how to solve such problems recently, if at all.

Design

We used a pretask-intervention-posttask design. The think-aloud method was used to capture students' thoughts during intervention. For the intervention, students were randomly assigned to one of two conditions, the compare and sequential condition, with 12 students in each condition (4 undergraduate students in the comparative group; 5 undergraduate students in the sequential group). Students in the compare condition were shown samples of worked solutions in pairs and asked to judge which of the two they felt was 'better'. Students in the sequential condition were shown the same set of worked solutions one-at-a-time and asked to assign each a score out of 5.

Instruments

Pre- & post-task: Students solved two tasks, a pre-task and a post-task. The pre-task required students to find the set of real numbers, x, such that $\frac{x+1}{x-7} > 3$, where $x \in \mathbb{R}\backslash\{7\}$. The post-task was similar in level of difficulty and required students to find the set of real numbers, x, such that $\frac{5x-2}{x+5} > 6$, where $x \in \mathbb{R}\backslash\{-5\}$. For both tasks, students were asked to write a solution as if submitting it for assessment.

Nine assessors (the research team and mathematics lecturers) ranked the pre- and post-tasks. This was done using ComPAIR which uses adaptive comparative judgement to strategically pair responses that have similar rank scores. Pairings are determined by considering each response's win/loss record from previous comparisons. As both pre- and post-tasks were ranked together, this meant that pairings could include a pre-task versus a pre-task, a pre-task versus a post-task, or a post-task versus a post-task. In total, assessors made 445 comparisons. This produced a ranking of students' work from 'best' (a score of 100) to worst (a score of 0).

Intervention: Students evaluated a set of eight worked solutions for the same pre-task problem. Solutions included (1) correct and incorrect answers which incorporated both minor errors and conceptual misunderstandings; (2) a range of methods and solution approaches from algebraic to graphical; (3) both high- and low-quality work; and (4) neat and messy solutions. All students were shown an identical set of eight solutions and in the same order for their respective experimental conditions.

Semi-structured interview: Students' pre- and post-tasks were placed side-by-side and students were asked to comment on any aspects of their solution they had chosen to keep the same or any they had changed and explain why.

PROCEDURE

All data collection occurred in a single problem-based interview lasting between 45 and 60 minutes. Students were given unlimited time to complete all activities. Students first completed the pre-task and did not receive feedback. Next, students evaluated the worked solutions, either in pairs or one-at-a-time. Students who compared solutions were asked to form a judgement for each pair based on three prompts: (1) quality of mathematical understanding; (2) quality of communication; and (3) overall quality. Students in the sequential group rated solutions on the same criteria and assigned scores out of five. Students were asked to think-out-loud while evaluating the worked solutions and were informed that there was no one correct evaluation strategy and that the way in which they either chose one solution as better or allocated their marks was up to them. Students were not provided with marking schemes or correct answers. Following evaluations, students completed the post-task without access to previous solutions. Finally, students compared their pre- and post-task solutions and commented on any changes made.

The think-aloud method (Ericsson & Simon, 1993) was used as the primary tool to access students' conscious thoughts during intervention. Interviews and think-aloud data were audio recorded and transcribed. Despite criticisms of think-aloud, notably its incompleteness in capturing underlying unconscious processes, the elicited data, though not exhaustive, remains valuable and informative about students' conscience cognitive processes.

ANALYSIS AND RESULTS

Knowledge gains from pre-task to post-task

Performance gains were measured by taking the difference between students' post-task and pre-task ranking scores. Data were screened for normality. Performance gains were normally distributed with skewness and kurtosis values within acceptable ranges; skewness ranged from -0.40 to 0.34, and kurtosis ranged from -0.22 to 0.83. The Levene's test of determining homogeneity of variance was not violated ($p = 0.578$). A two-sample *t*-test indicated significant differences in performance gains between groups, with students in the compare group (M = 22.1, SD = 16.7) found to have greater performance gains than those in the sequential group (M = 8.2, SD = 12.3); $t(22) = 2.32$, $p = 0.03$, $d = 0.95$.

Number of comparisons

Using students' think-aloud utterances, we examined the relationship between students' post-test outcomes and the number of comparisons made during intervention. A comparison was considered to be an instance where a student directly compared the

characteristics of one solution to another solution (e.g., "They both communicated their reasoning well" or "This one used a quicker method").

Students in the compare group made a total of 364 comparisons as opposed to 164 comparisons made by those in the sequential group. On average, students in the compare group (M=30.3, SD=12.8) made more comparisons than students in the one-at-a-time condition (M=13.7, SD=13.2), which was statistically significant $t(21) = 3.14$, $p = 0.005$, $d = 1.3$.

We questioned whether the act of making a comparison might be one reason for improved performance outcomes and examined the relationship between students' performance gains and the number of comparisons students generated. To explore this, a general linear model was used with the number of comparisons as a predictor variable and condition as a factor. Making more comparisons was not found to be predictive of performance gains, $F(1, 23) = 0.07$, $p = 0.790$, $\eta^2 < 0.01$. In short, even though students in the compare group outperformed those in the sequential group, it is unlikely to be because they made more comparisons.

Changes between pre- and post-task

Changes between pre- and post-tasks, as judged by the research team, were analysed to explore differences between groups. When analysing changes in students' work, the focus was not on the *quality* of students' changes, but rather on identifying *whether* a change had been made. Thematic analysis techniques were used to group instances of comparisons into codes. The data generated three categories: (1) Accuracy: making an improved attempt at solving the problem; (2) Communication: changing the amount of written explanation, including a heading, adding or removing visual components, changes in the choice of set notation; and (3) Method: changes to the choice of method such as changing from an algebraic to a graphical approach. Subjective qualities such as neatness, quality of the explanation, or whether one method was better than another were not included as the intent was to identify instances of change rather than assess whether these changes resulted in improvement in students' work.

As an example, Figure 1 shows the pre- and post-tasks for a student in Year 11. At pre-task, this student had found only the partial solution for the rational inequality, did not include any words or written explanation, and included algebraic manipulations only. At post-task, they had included the written annotations "If denominator is positive" and "If denominator is negative" as well as two number lines. This was counted as two changes under the category of communication. Additionally, the student changed their algebraic procedure by attempting to consider when the denominator might be positive or negative. Although the final answer was not correct, this was regarded as an improvement in understanding as the student showed awareness of the need to consider when the denominator is positive or negative. This was counted as one instance under the category of accuracy.

Results for the number of changes identified by the research team are displayed in Table 1. Students shown solutions in pairs made 34 changes between pre- and post-

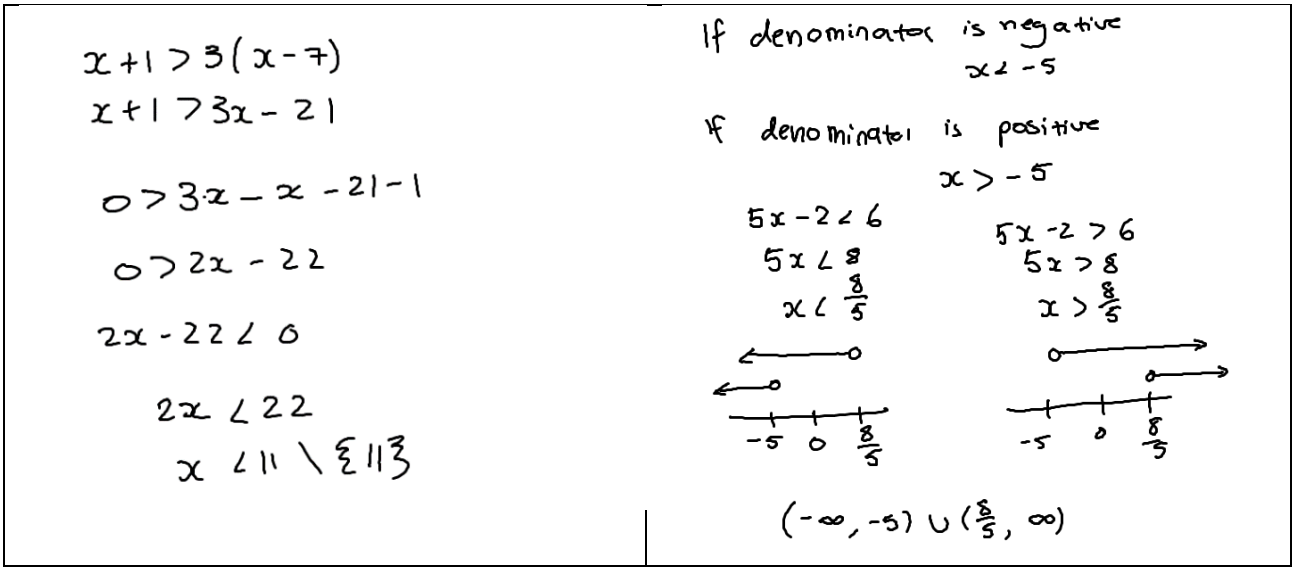

Figure 1: Sample pre-task (left) and post-task (right) completed by a Year 11 student

tasks while students shown solutions one-at-a-time made 13 changes. A Kruskal-Wallis test indicated this difference was significant, $H(1) = 6.60$, $p = 0.010$. For those in the compare group, most changes were to do with communication.

Next, the types of changes between pre- and post-task students reported during the interview stage were investigated. These comments were categorised using thematic analysis techniques similar to those described above. Elements students reported retaining/changing were grouped under the following five categories: (1) Accuracy: getting the final answer correct; (2) Communication: comments regarding the layout, amount of writing, headings, use of columns, etc.; (3) Method: changes in the approach used; (4) Presentation: making their solution neater; and (5) No changes made. The data are summarised in Table 2.

Elements	Experimental group	
	Sequential	Compare
Accuracy	6	10
Communication	5	23
Method	2	1
No meaningful change	4	-

Table 1 No. of changes made between pre- and post-task as judged by research team

A Kruskal-Wallis test indicated no statistically significant differences between conditions and the number of elements students reported retaining/changing, $H(1) = 2.71$, $p = 0.100$. While no effect was detected, we note that those in the compare group commented on more elements to do with how information was communicated than other categories, and that these students noticed more elements overall than those in the sequential group.

Elements	Experimental group	
	Sequential	Compare
Accuracy	7	6
Communication	4	11
Method	1	5
Presentation	1	1
No meaningful change	2	1
Total changes	13	23

Table 2 No. of elements retained/changed between pre- and post-task as reported by students

DISCUSSION

The present study investigated the impact of evaluating peer work comparatively on performance outcomes and sought to examine underlying factors contributing to any observed positive effects. Results indicate that students who evaluated peer solutions comparatively experienced greater performance gains than students who evaluated the same peer work one-at-a-time. It has long been known that comparing worked examples prepared by educators in mathematics classroom is beneficial for learning (Star & Rittle-Johnson, 2009). This paper shows that evaluating peer work comparatively is also useful, and that comparative judgement appears effective in the context of mathematics.

Why comparing solutions was more effective than presenting the same solutions individually is still not clear. Existing literature emphasises the benefits of comparing in learning. It is plausible that the improved performance outcomes when comparing is because students in the compare group were explicitly instructed to compare solutions and, as such, generated more comparisons, suggesting that these students had more opportunities to discern structural similarities and differences across solutions. However, we found no significant relationship between the number of comparisons made and performance gain, consistent with previous research (Star & Rittle-Johnson, 2009). Hence the underlying factor is unlikely to be the *quantity* of comparisons students generate, but rather, the substance and quality of these comparisons. Further research is needed to investigate the nature of these comparisons and their relationship with learning.

Furthermore, comparative judgement appears to influence the types of changes students made to their own work. Students who compared solutions were more likely to modify how their work was communicated, made more changes overall to their work, and were better at verbalising these changes. In contrast, students evaluating worked solutions individually did not make as many changes to their own work and were less likely to verbalise these changes. This supports the claim made by Kimbell

(2020) that the reason comparative judgement results in improved performance outcomes is because it enhances students' ability to articulate elements that constitute high-quality work. For the current study, students who engaged in comparison were more adept at identifying and expressing changes in the overall quality of their work with less emphasis on the correctness of their final answer. The implication for educators is that comparative judgement may be a useful tool when wanting to direct students' attention beyond simply solving a problem correctly and towards more holistic elements of quality.

LIMITATIONS

The current study has some limitations. First, it was conducted in a laboratory setting which may have magnified the demonstrated effects compared to a naturalistic classroom environment (Alfieri et al., 2013). Future researchers may wish to see whether the findings from this study are replicated where comparative judgement is used as an authentic classroom activity. Second, this study included a small sample size. Statistical analysis should be interpreted with caution as this may have weakened findings. Future research with larger cohorts is warranted to validate and extend these findings.

REFERENCES

Alfieri, L., Nokes-Malach, T. J., & Schunn, C. D. (2013). Learning Through Case Comparisons: A Meta-Analytic Review. *Educational Psychologist, 48*(2), 87–113. https://doi.org/10.1080/00461520.2013.775712

Bartholomew, S. R., Mentzer, N., Jones, M., Sherman, D., & Baniya, S. (2022). Learning by evaluating (LbE) through adaptive comparative judgment. *International Journal of Technology and Design Education*, *32*(2), 1191–1205. https://doi.org/10.1007/s10798-020-09639-1

Bartholomew, S. R., Strimel, G. J., & Yoshikawa, E. (2019). Using adaptive comparative judgment for student formative feedback and learning during a middle school design project. *International Journal of Technology and Design Education*, *29*, 363–385. https://doi.org/10.1007/s10798-018-9442-7

Begolli, K. N., & Richland, L. E. (2016). Teaching mathematics by comparison: Analog visibility as a double-edged sword. *Journal of Educational Psychology*, *108*(2), 194–213. https://doi.org/10.1037/edu0000056

Boaler, J. (1998). Open and closed mathematics: Student experiences and understandings. *Journal for Research in Mathematics Education*, *29*(1), 41–62.

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*, 86. https://doi.org/10.3389/feduc.2018.00086

Ericsson, K. A., & Simon, H. A. (1993). *Protocol Analysis: Verbal reports as data (revised edition)*. Bradford Books/MIT Press.

Große, C. S. (2014). Mathematics learning with multiple solution methods: Effects of types of solutions and learners' activity. *Instructional Science*, *42*(5), 715–745. https://doi.org/10.1007/s11251-014-9312-y

Kimbell, R. (2020). Capability, Quality and Judgement: Learners' Experiences of Assessment. In P. Williams & D. Barlex (Eds.), *Pedagogy for Technology Education in Secondary Schools. Contemporary Issues in Technology Education* (pp. 201–217). Springer. https://doi.org/10.1007/978-3-030-41548-8_11

Marton, F. (2015). *Necessary conditions of learning*. Routledge.

Seery, N., & Canty, D. (2018). Assessment and Learning: The Proximal and Distal Effects of Comparative Judgment. In M. J. De Vries (Ed.), *Handbook of Technology Education* (pp. 735–748). Springer International Publishing. https://doi.org/10.1007/978-3-319-44687-5_54

Star, J. R., & Rittle-Johnson, B. (2009). It pays to compare: An experimental study on computational estimation. *Journal of Experimental Child Psychology*, *102*(4), 408–426. https://doi.org/10.1016/j.jecp.2008.11.004