AN ASSESSMENT OF PEER INSTRUCTION IN LARGE FIRST YEAR MATHEMATICS COURSES

RAYMOND VOZZO, STUART JOHNSON, AND JONATHAN TUKE

Abstract. Many recent studies have expounded the benefits of active learning in tertiary education. It can be challenging to implement these techniques at large scale (for example in first year mathematics courses). A common method for actively engaging students in large classes is through online quizzes, which may include peer instruction. In this paper, we investigate the effect of having students answer quiz-style questions during class both with and without discussion in a first year mathematics course. We also investigate the short- and long-term effects of each protocol.

We find that peer instruction improves student learning in mathematics in the following ways: First, when the responses to questions was measured before and after peer instruction the proportion of questions answered correctly increased by 0.2; second, when correct responses were compared to similar questions the following week the proportion correct increased by 0.34 (compared to 0.07 for the control); finally, when measured at the end of the semester the proportion of questions answered correctly increased by 0.42 (compared to 0.2 for the control).

Contents

1. Introduction

Active learning has been proposed as a method for improving learning and outcomes in STEM areas (see for example [5]). For large classes many of the most popular methods for introducing active learning involve so-called clickers [4], which involve students using technology to answer quiz questions. This style of active learning may also include peer instruction, in the sense originally due to Mazur [3], which requires students to answer a question, discuss with their fellow students and then answer again. Methods that can be employed effectively in large classes (for example, more than 400 students) are particularly important in mathematics as many science and engineering faculties have large mathematics service courses in first year undergraduate degrees.

Following the work of Mazur in employing peer instruction in physics, it has since been adapted to many other areas [10]. In mathematics, studies related to the use of peer instruction have considered various aspects including: the effectiveness against traditional workshops in linear algebra [9] and in calculus [6] and the problem of improving questions to achieve optimal effectiveness [2, 8, 11].

In this article, we study the effectiveness of peer instruction in undergraduate mathematics (both in linear algebra and in calculus) by comparing student performance in answering questions during class with or without structured discussion. This allows us to measure the value of augmenting quiz-based active learning in large mathematics with peer interactions.

2. Methodology

2.1. **Background.** This study was conducted in a first year undergraduate mathematics course containing parallel streams of calculus and linear algebra, with a diverse cohort including mathematics, computer science, engineering and science students. Enrolments for the semester were approximately 550 students. For most students this is their second semester of university mathematics. The course utilises a flipped classroom model, with videos and notes on the course material available online in the learning management system; a one hour workshop each week for the entire class (in a very large workshop theatre) where the students participate in quizzes, with follow up discussion from the workshopr where required; and weekly tutorials, where students work in groups of 4 or 5 at whiteboards solving problems. There are several components of assessment throughout the semester (including weekly written and online assignments, and an invigilated test) and a final written exam.

In the workshop, peer instruction is often used, where a question is asked and students provide answers (using their phones or other devices to access Mentimeter[1]), and are then invited to discuss the question with their peers and potentially change answers.

For the study, each question asked was randomly allocated one of two treatments:

- **control**, in which the question was posed and students have only one opportunity to answer, after which the solution would be revealed and explained if necessary.

[1] www.mentimeter.com

No particular attempt was made either to force students to discuss with their peers or to not discuss at all, they were allowed to answer in a natural way;

- **discussion**, in which students were instructed to give an initial answer without any interactions with their peers, with the collective answers shown to the class before giving the students a chance to discuss their answers, and change them if they wish. Then the solutions are shown.

The questions asked were mostly conceptual in nature, highlighting fundamental aspects of the material or particular topics that students typically have difficulty with. Generally it was expected that students could answer questions without the need for a great deal of calculation. Some examples of questions used can be found in Appendix A.

The effect of peer instruction was measured in a number of different ways. In cases where students are asked to discuss questions and then given the opportunity to change their answer we can compare the initial responses with those obtained after the discussion. To obtain a measure of the effect of peer instruction against a control group we measured responses over consecutive weeks. In each weekly class, students are first asked new questions about the same concepts as the questions from the previous week. These are designed to test conceptual understanding rather than recall of answers from the previous week (see Appendix A for examples). This allows us to measure the effect of peer instruction by comparing performance on these questions for each of the "control" and "discussion" treatments.

In addition, some questions were repeated at the end of the semester in a revision class, allowing longer term retention of knowledge to be measured.

2.2. **Statistical methodology.** Each week, four new questions were asked in the quiz. These consisted of two algebra questions, and two calculus questions. Within each area (calculus or algebra), one question was a control question, and one was a discussion question. The allocation to either control or discussion was done by randomization.

In weeks 2-11 four additional questions were asked, these were related to the questions from the previous week. In week 12, there were 15 questions that were all related to questions from previous weeks.

The form of the experiment is based on a matched pairs design, in that we have a pair of questions that were discussed in the previous week, and also a pair of question that were not discussed and acted as the control. All of the analysis was performed in R [7] using the program RStudio[2]. We fitted four models to address the effect of repetition on the retention rates of students. In each case, we fitted a generalised linear mixed-effects model (GLMM) with the outcome variable being the number of correct answers, and the predictors the offering and whether that the question had been reviewed—where appropriate. As well, to account for the repeated measurements, we included a random intercept for each question. The GLMM were fitted using the `lme4` package [1]. For each model, we then predicted the probability of getting a question correct as given in Tables 1, 2, 3 and 4.

[2]https://posit.co/download/rstudio-desktop/

3. Results

Figure 1 gives an overview of the experimental design, and indicates when questions were repeated. Some questions are the same question (brown triangle), while some are different questions, but the same concept (blue circle). The offering on the x-axis gives the week and the order, so for example 02-3 is the third question in the second week. The ID identifies each question. The first part gives the week, the letter indicates the subject (A is Algebra, while C is Calculus) and the number indicates whether there is more than one question in a subject, so W02A3 is the third question in Algebra for Week 2. We see that some questions were repeated once, for example W01A1, while some may be repeated up to four times (e.g. W04C2).

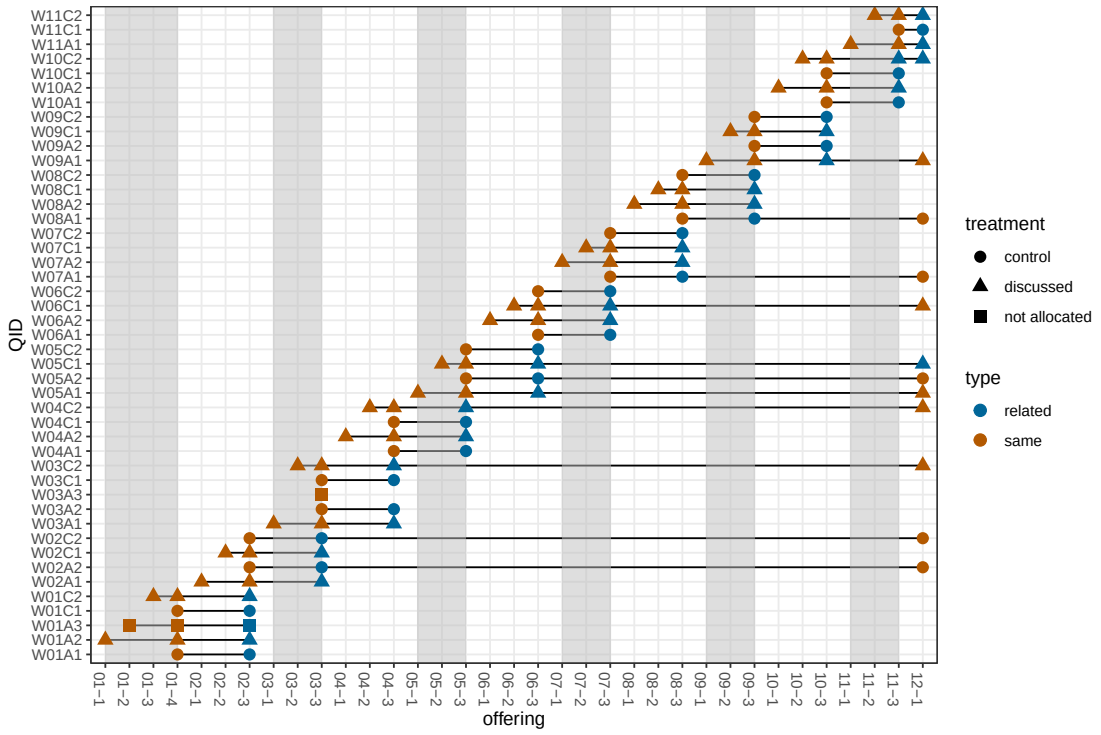


Figure 1. Figure of experimental design showing when questions were repeated. Some questions are the same question (brown triangle), while some are different questions, but the same concept (blue circle). The offering on the x-axis gives the week and the order, the ID identifies each question.

Figure 2 shows the proportion correct for each attempt of a question. The thick lines show the average proportion.

In more detail, we fitted four seperate models:

- Repeat Model, which compares the first and second attempt at discussion questions in the same workshop.

- Short-term Model A, which compares the first attempt at control and discussion questions with the attempt at the related questions in the following week. Note that it is the first attempt at the discussion question being used.
- Short-term Model B, which compares the control and the second attempt at the discussion questions with the attempt at the related questions in the following week. Note that it is the second attempt at the discussion question being used.
- Long-term Model, which compares the first attempt at control and discussion questions with the attempt in Week 12.

The results of each of these models is shown in Tables 1, 2, 3 and 4, respectively.

First we see that over time there is an improvement in understanding as seen by an increased proportion of correct answers for second attempts and beyond. This is also seen in Tables 1, 2, 3 and 4, where in all cases there is a positive improvement in proportion correct. Also we see that the increase in understanding from the first attempt is much larger for discussion questions compared to the control questions (Table 4). This may be that we can see that the proportion correct for discussion questions on the first attempt is much lower than the control questions. As the questions were randomly allocated to control or discussion, this is surprising—see Section 4 for a discussion on the authors' thought on why this occurred.

Treatment	1st	2nd	Improvement
Discussed	0.42	0.63	0.2

TABLE 1. Repeat Model. Proportion correct for questions at the first and then second attempt in the same workshop.

Treatment	First week	Second week	Improvement
Control	0.67	0.74	0.07
Discussion	0.43	0.77	0.34

TABLE 2. Short-term Model A. Proportion correct for questions at the first and then second attempt in the next workshop. Note that we are using the first attempt for the discussed questions in the first week.

Table 5 contains the estimates and P-value for each of the models that we considered. We see a statistically significant effect (5% level) of offering and treatment (control versus discussion). Also we see that there is a statistically significant interaction of offering and treatment for the Short-term and Long-term models. As the interpretation of two-way interactions are difficult to interpret, we advise looking at the predicted probabilities of getting the answers correct as given in Tables 1, 2, 3 and 4.

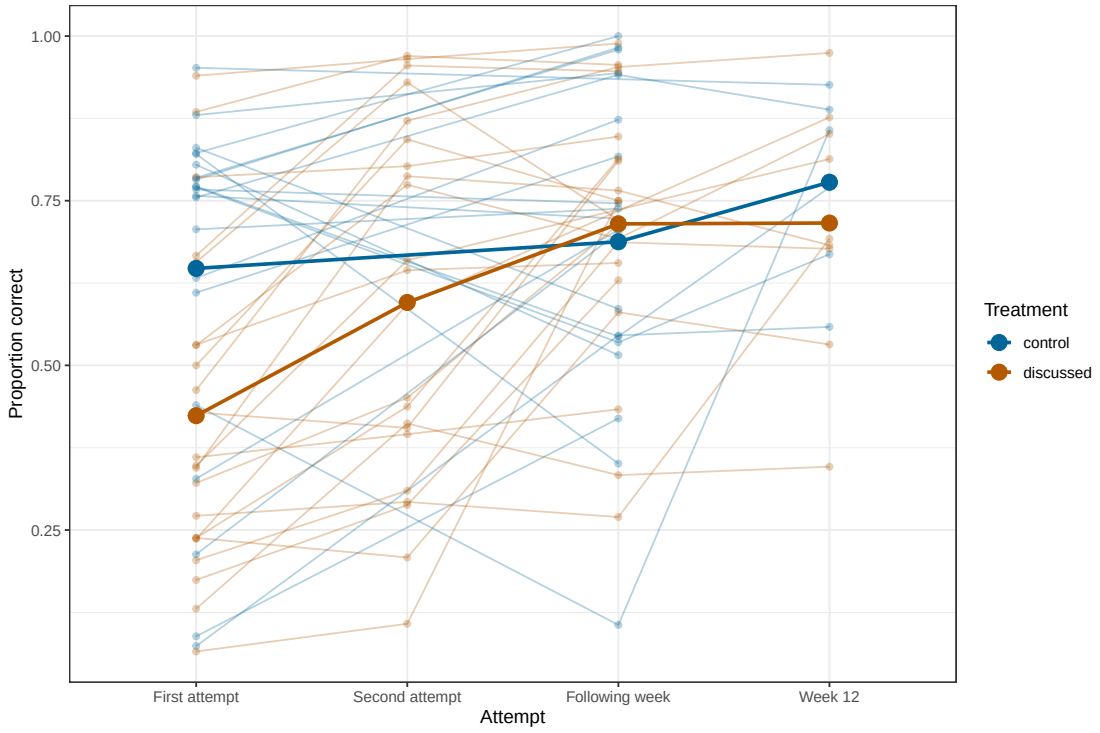


FIGURE 2. Proportion answered correctly for each attempt of the questions. The colour indicates if the question is a discussion or a control question. The lines correspond to a single question.

Treatment	First week	Second week	Improvement
Control	0.67	0.74	0.07
Discussion	0.64	0.79	0.15

TABLE 3. Short-term Model B. Proportion correct for questions at the first and then second attempt in the next workshop. Note that we are using the second attempt for the discussed questions in the first week.

Treatment	First offering	Week 12	Improvement
Control	0.55	0.75	0.2
Discussion	0.28	0.70	0.42

TABLE 4. Long-term Model. Proportion correct for questions at the first and then second attempt in the Week 12 workshop.

Model	Predictor	Estimate	P-value
Repeat	Repeat	0.83	1.8×10^{-33}
Short-term Model A	Discussion	-0.99	3.01×10^{-3}
	Second week	0.34	1.05×10^{-6}
	Interaction	1.16	1.11×10^{-29}
Short-term Model B	Discussion	-0.13	7.25×10^{-1}
	Second week	0.35	1.04×10^{-6}
	Interaction	0.42	4.57×10^{-5}
Long-term	Control	-1.13	2.34×10^{-2}
	Week 12	0.91	5.97×10^{-11}
	Interaction	0.88	7.84×10^{-6}

TABLE 5. Coefficients and P-values for each of the three models fitted to look at the effect of offering and discussion on the probability of getting a question correct.

4. Limitations and outlook

Here we outline some of the limitations of this project and possible future directions for study. We also offer an explanation as to the interesting difference noted from Figure 2 between the proportion of correct answers on the initial attempt for the control questions and the discussion questions.

The main limitation was the fact that the discussion environment was largely determined by how students chose to engage. The class was in a large workshop theatre and it was up to students to choose to sit with others with whom they could discuss the questions, so some students may not have engaged in discussion. It is possible that if they changed their answer after the discussion period it may have simply been the result of seeing which answer the majority chose, rather than the outcome of a reasoned discussion.

Furthermore, this also meant that the control and repeated questions were not strictly discussion free, which could explain the difference in success of the first attempts at questions. It is possible that during the discussion questions, students were explicitly instructed (and monitored) to *not* discuss the questions on their first attempt; whereas during the control questions this was not done, so some discussion naturally occurred during this time. This actually represents a more natural classroom situation than the somewhat artificial enforcement of silence during a question. Thus the "control" questions can be viewed as "non-peer-instruction" questions, while the discussion questions are, as previously described, structured peer instruction questions.

This observation prompted the comparison of the (first and only) attempt at the control questions with the *second*, post-discussion attempt at the discussion questions, which is the model labelled Short-term Model B, shown in table 3. This model shows that there is still an improvement in answering the related questions the following week, suggesting

that having a genuine solo attempt at a question before discussion resulted in improved understanding.

Further study is warranted here, with a potential improvement being to ensure that the review questions that, which are asked the following week, have a "no discussion" policy while being answered. This would ensure that the short term improvements described in Table 2 are accurate. It would also be interesting to repeat this with different classroom setups; for example, in smaller rooms, or flat-floor rooms that are more amenable to collaborative discussions than a large workshop theatre.

5. Conclusion

The effect of peer instruction was measured in three ways. First, by comparing the responses of students to the same question immediately after discussion with peers (Repeat model). Second, by asking related questions the following week and measuring performance in comparison with control questions for which no peer instruction was used (Short-term models). Finally by repeating some questions in a revision session at the end of the semester (Long-term model).

In response to the significant difference in first attempts at control and discussion questions, the Short-term models were split into two versions, comparing either the first or second attempt at a discussion question with the attempt at a related question the following week, to account for the possibility that the control might involve student discussion and hence be more comparable with the post discussion question.

The results show a positive effect of peer instructions in all four measures.

Specifically, using the first measure we find that peer discussion improves the proportion of correct responses by 0.2. Using the second measure, we find that while the proportion correct for the control questions increased by 0.07, for the peer instruction questions this improvement was 0.34 in model A or 0.15 in model B. Finally, at the end of the semester, while there was improvement for both sets of questions, for the control questions this was 0.2 while for the peer instruction questions, this was 0.42.

Acknowledgments

The authors would like to thank Nickolas Falkner for helpful feedback on the interpretation of the results and Tanya Evans for helpful comments on the first draft.

Appendix A. Examples of questions

We provide some examples of questions used during the study that demonstrate the sort of thing that students were typically asked. The examples below show both conceptual questions and computational questions. Note that if any computations were required they were both fairly simple and designed to test understanding of a particular technique or result. In each case two versions of a question are presented, the initial one, and the related question which was asked the following week. The related questions are designed to test the same concept, and to require genuine understanding of the concept, rather than being able to be correctly answered purely by knowing the answer to the original question.

Example. Which of the following are linear transformations?

A. $F(x, y) = (x, y)$

B. $F(x, y) = (0, 0)$

C. $F(x, y) = (1, 1)$

D. $F(x, y) = (xy, 0)$

E. $F(x, y) = (y, y)$.

Related question. Which of the following are linear transformations?

A. $F(x, y) = (y, x)$

B. $F(x, y) = (0, 1)$

C. $F(x, y) = (x + y, 0)$

D. $F(x, y) = (xy, xy)$

E. $F(x, y) = (y, y + 1)$.

Example. Suppose $A = \begin{bmatrix} 2 & * \\ * & 3 \end{bmatrix}$ is not invertible. What are the eigenvalues of A?

A. 0

B. 2

C. 3

D. 5

E. 6

F. There is not enough information to determine this

Related question. Suppose $A = \begin{bmatrix} -1 & * \\ * & 2 \end{bmatrix}$ is not invertible. What are the eigenvalues of A?

A. -2

B. -1

C. 0

D. 1

E. 3

F. There is not enough information to determine this

Example. Fill in the blank.

$\sum_{n=1}^{\infty} a_n$ converges _____ $\lim_{n\to\infty} a_n = 0$.

A. $\Rightarrow$

B. $\Leftarrow$

C. $\Leftrightarrow$

Related question. Fill in the blank.

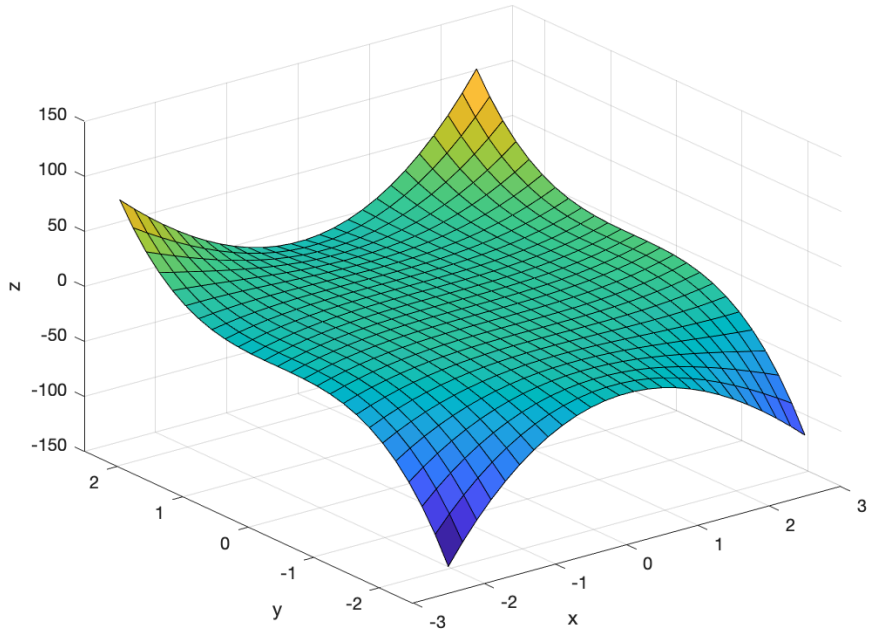
Suppose that $a_n > 0$ is a decreasing sequence.
Then $\sum_{n=1}^{\infty} (-1)^n a_n$ converges _____ $\lim_{n\to\infty} a_n = 0$.

A. $\Rightarrow$

B. $\Leftarrow$

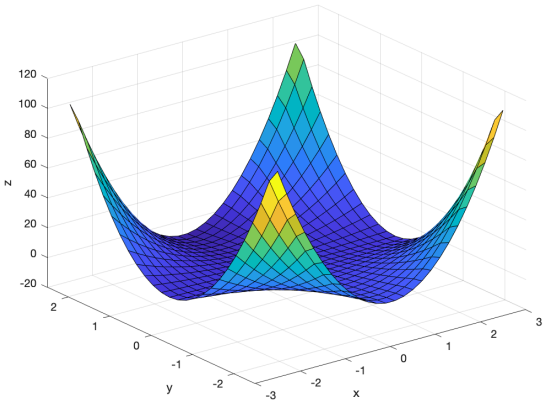
C. $\Leftrightarrow$

Example. Consider the graph of the function $f(x, y)$ below.

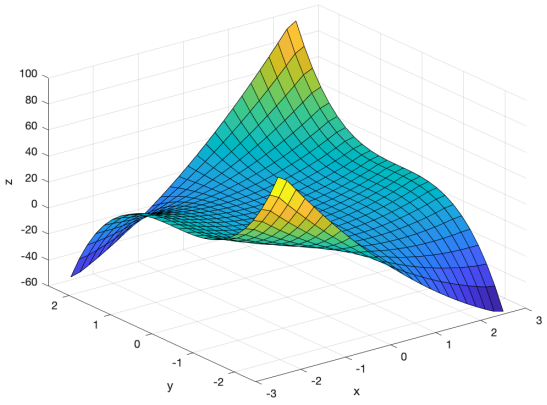


Which of the following is f_x and which is f_y?

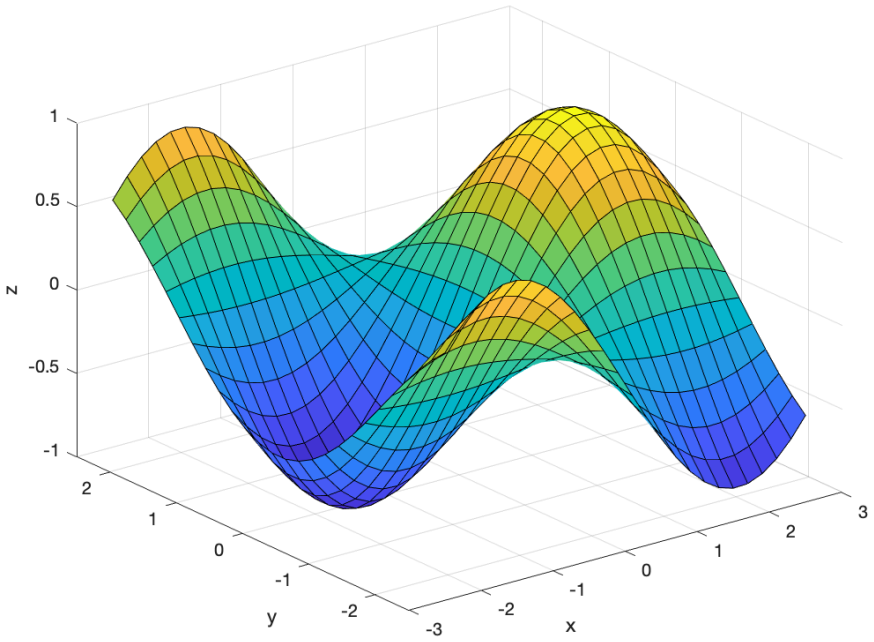
A.



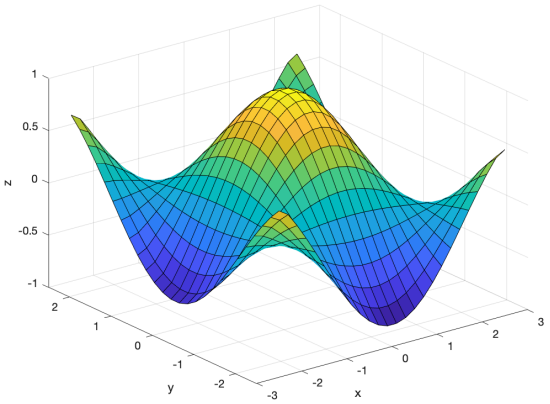
B.

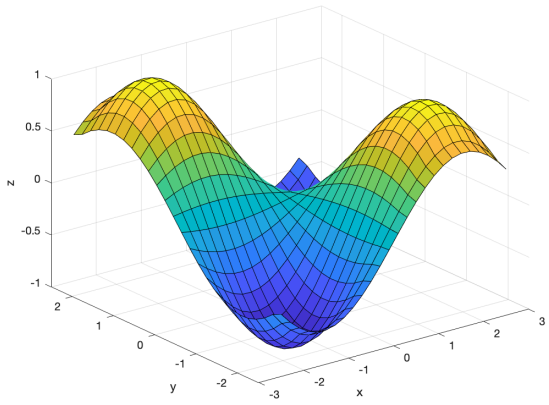


Related question. Consider the graph of the function $f(x, y)$ below.



Which of the following is f_x and which is f_y?

A. 

B. 

References

[1] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015.

[2] Kelly Cline, Holly Zullo, Jonathan Duncan, Ann Stewart, and Marie Snipes. Creating discussions with classroom voting in linear algebra. *International Journal of Mathematical Education in Science and Technology*, 44(8):1131–1142, 2013.

[3] Catherine H. Crouch and Eric Mazur. Peer Instruction: Ten years of experience and results. *American Journal of Physics*, 69(9):970–977, 09 2001.

[4] Douglas Duncan and Eric Mazur. *Clickers in the Classroom: How to Enhance Science Teaching Using Classroom Response Systems*. Pearson Series in Educational Innovation: Instructor Resources for Physics Series. Pearson Education, 2005.

[5] Scott Freeman, Sarah L. Eddy, Miles McDonough, Michelle K. Smith, Nnadozie Okoroafor, Hannah Jordt, and Mary Pat Wenderoth. Active learning increases student performance in science, engineering, and mathematics. *Proc. Natl. Acad. Sci. USA*, 111(23):8410–8415, 2014.

[6] Scott Pilzer. Peer instruction in physics and mathematics. *PRIMUS*, 11(2):185–192, 2001.

[7] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2023.

[8] Everilis Santana-Vega Robyn L. Miller and Maria S. Terrell. Can good questions and peer discussion improve calculus instruction? *PRIMUS*, 16(3):193–203, 2006.

[9] Katiuscia Costa Barros Teixeira. Pedagogical strategies to enhance learning in a linear algebra course. *PRIMUS*, 33(2):152–174, 2023.

[10] Jonathan G. Tullis and Robert L. Goldstone. Why does peer instruction benefit student learning? *Cognitive Research: Principles and Implications*, 5(1):15, 2020.

[11] Hidetaka Yamaoka, Makoto Nishi, Tetsuya Taniguchi, and Tomoshige Kudo. Practice of calculus lecture using peer instruction by audience response system. In *Proceedings of the 2020 11th International Conference on E-Education, E-Business, E-Management, and E-Learning*, IC4E '20, page 279–283, New York, NY, USA, 2020. Association for Computing Machinery.

(Raymond Vozzo) School of Computer and Mathematical Sciences, University of Adelaide, Adelaide, SA 5005, Australia

Email address: raymond.vozzo@adelaide.edu.au

(Stuart Johnson) School of Computer and Mathematical Sciences, University of Adelaide, Adelaide, SA 5005, Australia

Email address: stuart.johnson@adelaide.edu.au

(Jonathan Tuke) School of Computer and Mathematical Sciences, University of Adelaide, Adelaide, SA 5005, Australia

Email address: simon.tuke@adelaide.edu.au