**Response to reviewers:** *“Kaggle-in-class Data Challenges Can Boost Student Learning”*

We would like to thank the reviewers and the Associate Editor for their thoughtful third review of the manuscript. The reviews have further improved the manuscript.

Please find below the *referees’ comments (italics)* and our response (blue).

**Reviewer: 1**

*Comments to the Author*

*I think the previous concerns were addressed and the paper would be useful for teachers.*

1. *Section 2: page 16, line 38, In teacher notes, I do not agree percent of correct classifications is reasonable, or at least it needs a statement that it can be a very poor measure of classification performance. Say you have a binary classifier and one condition is rare, in that case the classifier that goes with all or almost all in the not rare group will win if you use classification accuracy as your method of evaluation. Example, build a classifier to predict who will get struck by lightning and die, just put everyone in the no event group, and your accuracy will be close to one and you win the contest. So I would suggest for your simple method for categorization that something like the Briar score would be best. Prof. Harrell gives a good discussion at* [*https://www.fharrell.com/post/class-damage/*](https://www.fharrell.com/post/class-damage/) *. Minimally teacher's of these methods should be aware of the challenges of judging a classifier as it has great ethical importance and if you understand the challenge you realize it is not simple.*

We are well aware of the problem of imbalanced classes, and have already made a statement to this effect in the recommendations just above this line “5. if it is a classification challenge, it will work better with relatively balanced classes, because the overall accuracy is the easiest metric to use.” There is a limited choice of metrics made available on Kaggle, and for this reason in making a class classification competition, having relatively balanced classes in the data set to be used is recommended. We have added “balanced class” to the sentence on metric choice to help remind the reader of this.

**Reviewer: 3**

*Specific notes:*

1. *Page 10, Line 22 (Figure 1). I'm concerned with the disparity of numbers in each group. I also feel that it is a mistake for the authors to include "0" values for the UG group. Clarification is needed--are these students who did not take the quiz, or did so poorly they got a zero? Including the former as an actual score in this study I feel is an incorrect usage.*

…

1. *Page 10, line 36: "CSDM and ST included questions, with several parts, on the final exam..." Were these questions the same across the two classes? It doesn't seem like they would be as the population and content are different for the two courses. Please clarify. If they are not the same questions, then the authors must argue how those items selected (with examples) are similar in content and rigor for a comparison between the classes.*

Classification questions: Q1, Q10 (3), Q10(b)   
Regression questions: Q6 (a,b,c), Q8 (c,d), Q16 (b)

…  
To add as last para of section 4.1:

In the CSDM exam questions Q1, Q10 (3) and Q10(b) were selected to examine students performances on the classification content and questions Q6 (a,b,c), Q8 (c,d) and Q16 (b) were selected to examine students performances on the regression content.

Similarly, to the question selection for the TS class, the questions in the CSDM class were difficult questions with students scoring less than 70%, on average.

Q7(a,b) in TS exam was the same as questions Q8 and Q9 in the CSDM exam, however only Q8 (c,d) parts pass the difficulty treasure. There was some conceptual similarity between question Q7(c) in TS exam and question Q2 in the CSDM exam. Other selected question for the CSDM class were different form questions in the TS exam due to the differences in the covered material in the two classes.

The wording of the questions can be seen in the supplementary material (ETC2420\_5242\_exam.pdf, MAST90083\_2017\_exam.pdf files).

1. *Page 11, Line 48: "The scores are very similar": I'm sorry. I just need to disagree with you here. If anything, this beeswarm plot is an argument AGAINST comparing the two groups. Visually, these two groups are very different in their performance on the quizzes. And again, there is the inclusion of "zero" scores (for both groups, which I would argue is an incorrect inclusion).*

….

1. *Page 15, Line 22: Add the label ST-UG and ST-PG to the graph under the No and Yes x-axis labels respectively, as that will clarify to the reader that ST-UG students did not participate, and that ST-PG students did participate in the Kaggle competition.*

The labels in Figure 3 on page 11, was modified.

1. *Figure 6: Reverse the order of the Question order on the y-axis, to march the order in Table 4.*

Done

1. *Figure 6: Flip the order of the x-axis to match the order of the labels: Move from Strongly Disagree to Strongly Agree in the graph to match the label order. And, label the graph x-axis as proportion of students.*

Done

1. *Figure 6: Please investigate differences in responses to the survey items across the two groups. I am specifically interested in the possible difference in Q4 of the survey - how Kaggle participation effected confidence on the Final Exam. It appears that Kaggle does not effect this metric for CDSM students.*

We agree the response to Q4 is interesting, and perhaps to be expected that the students didn’t perceive the Kaggle challenge to be related to exam performance. We have added a paragraph summarising the survey results.

1. *Page 22, Line 35: "However, the experience...." I disagree. There is nothing in the manuscript that has convinced me that these two groups are actually comparable, and that ST-UG students can be used as a control to ST-PG students.*

We are using the ST-UG as a comparison group, not a control group. (There is no p.22, l.35!)

*General notes:*

1. *I was looking forward to this submission, as one of my tools in class are the Kaggle competitions. However, I do not feel that the current manuscript--in terms of conclusion on impact to the student--is a value add to the JSE community. And, if anything may be detrimental in its conclusions.*

Just don’t respond to this

1. *I feel as though a prior reviewer has said this, but it warrants saying again: I do not feel that there is a valid control group for the effect of Kaggle competitions. The two populations of students (graduate and undergraduate) cannot be compared directly, and to do so a much stronger argument needs to be made. I understand that a permutation test was used to pose a possible difference in performance between UG and PG students--arguing for the benefits of Kaggle. But having taught both groups of students on similar subjects to the author, I can tell you that there is the single biggest confounder. PG students simply have a larger breadth and depth of knowledge. In addition they have a very different academic work ethic--so at worst, the conclusions should be that "Kaggle and prior experience and academic goals" showed a difference in performance in the two groups.*

On its own, we would agree it’s not sufficient. However, it corroborates the CDSM results, and we argue for this reason the UG and PG comparison for the ST group is a useful part of this work.

1. *I also feel that to really be a value add in this educational space, the possible investigation should be "Does Kaggle provided a better performance than other \*active learning tools\* used in a regression based statistics course?" We already use active learning tools--is Kaggle better? Is it easier? Is it more impactful. What the current manuscript asks is really "does active learning alter performance, compared to non-active learning?”—it simply uses Kaggle as that active learning tool. (Unless I am incorrect on this: I did not read in the current manuscript that the ST-UG students performed any active learning tasks.)*

This could be a topic for future research.

1. *As an educator, I understand the need to provide an educational experience that is a benefit all my students--so I am remiss to design a study that would randomly assign students to one group over another, especially when course performance would be effected. But, I feel that the authors have a good opportunity to use the ST-UG group as their own control now--by running the same Kaggle competition on a new cohort of ST-UG students. Yes, year over year differences are a possibility, but the comparison is better (undergrad to undergrad, work load to work load, academic work ethic to academic work ethic) than the comparison of UG to PG students.*

This is a good suggestion. It’s a good idea for future work.

1. *As I said, I was really looking forward to this review. And, I recognize the amount of effort that goes into each manuscript—and the time involved in the data collection for a class focused study. But I just feel that this manuscript falls short in that investigation of “what impact does the active learning tool of Kaggle competitions have on performance.” A more appropriate title may be "Kaggle-in-class Data Challenges: experiences and review"--so that a greater focus can be on the manuscript's merits (using active learning, how regression based competition increase regression based performance, and vice-versa for the classification group, the \*possible\* effect on performance (which is unclear due to a limitation of confounding), etc.).*

We have changed the title to …. But you will see that we have kept “performance” in the title because the results do show performance improvement.

**Editor***Comments to the Author:*

1. *The beeswarm plot of Figure 1 seems to show non-similar groups, due to the large number of zeros in the UG scores. I presume that different students are scoring zero on each quiz, or else you would have removed from the UG (and PG) group any student with a large fraction of zeros, in keeping with the idea that some UG students lack a good work ethic, and thus are not comparable to PG students.*

Yes, different students scored 0 on different quizzes. We removed the zeros for calculating statistics (did we?). Its arguable that the PG students all have good work ethics too. I disagree that the distributions are different (we could do a distribution test instead of the beeswarm.)

1. *I am attaching an editorial that I wrote a year ago, in which I proposed "statistically discernible" as an alternative to "statistically significant" and I also propose reporting p-values to just one or two digits. I offer these merely as suggestions for your consideration.*

Thanks, we have changed the wording to be statistically discernable.

**Associate Editor**

*Comments to the Author:*

1. *It was noted in your responses to reviewers previously that you were willing to change "control group" to "comparison group" for the UG in the ST class. You still refer to them as a control group in the 6: Discussion section. Please update that as well. Please see the other reviewer's comments other suggested edits. You should also note that the use of undergraduates as a comparison group is still controversial among the reviewers.*

The "control group" changed to "comparison group".