On the Effect of Question Ordering on Performance and Confidence in Computer Science Examinations

Brian Harrington
Dept. of Computer and
Mathematical Sciences
University of Toronto Scarborough
brian.harrington@utsc.utoronto.ca

Jingyiran Li
Dept. of Computer and
Mathematical Sciences
University of Toronto Scarborough
jingyiran.li@mail.utoronto.ca

Mohamed Moustafa
Dept. of Computer and
Mathematical Sciences
University of Toronto Scarborough
mohammad.moustafa@mail.utoronto.ca

Marzieh Ahmadzadeh
Dept. of Computer and
Mathematical Sciences
University of Toronto Scarborough
marzieh.ahmadzadeh@utoronto.ca

Nick Cheng
Dept. of Computer and
Mathematical Sciences
University of Toronto Scarborough
nick@utsc.utoronto.ca

ABSTRACT

Most computer science examinations tend to start with the easiest questions and progress towards the more difficult material. Whether this is because of the highly scaffolded nature of the course, an attempt to 'ease students in', or simply by convention, is unclear. However, there is a great deal of data from the psychology literature to suggest that human perception of the difficulty or discomfort of a task is disproportionately affected by the last part of the task completed. Therefore, is it possible that by structuring our exams in an easy-to-hard fashion, we are causing students to perceive the test as more difficult than it actually is? Could changing the question order allow us to change students' perception of their own achievement? What effect could this have on actual performance? This paper attempts to answer these questions by randomly assigning students to write exams ordered either easy-to-hard (referred as 'Easy-Difficult') or hard-to-easy ('Difficult-Easy), then ask them to predict their marks on per-question basis. We find that the question ordering has a small but not statistically significant effect on the performance, and virtually no effect on predicted marks when treating the entire class as one unstratified sample. However, the effect was significant for certain subgroups created via stratification. In particular, swapping the order of the questions may have hurt the performance of international students, but significantly raised both the performance and confidence of female students.

CCS CONCEPTS

• **Social and professional topics → Computing education**; **Computer science education**; **CS1**;

KEYWORDS

CS1, examination, question ordering, confidence, gender, international students

ACM Reference Format:
Brian Harrington, Jingyiran Li, Mohamed Moustafa, Marzieh Ahmadzadeh, and Nick Cheng. 2019. On the Effect of Question Ordering on Performance and Confidence in Computer Science Examinations. In *Proceedings of the 50th ACM Technical Symposium on Computer Science Education (SIGCSE '19), February 27-March 2, 2019, Minneapolis, MN, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3287324.3287398

1 INTRODUCTION

When creating final examinations for courses, the natural tendency is to either include questions in the order in which they were covered in class, or to start with the easiest questions and work towards the most difficult ones. Quite often these two orderings are the same. However, there is much research in the psychology and education literature that shows us that human perception of the difficulty or unpleasantness of a task follows the 'peak-end' rule rule[8]. This rule says that a subject's perception is generally estimated as the mean between the most extreme and last remembered experiences in a task. Translating this idea to examinations implies that students perceptions of the difficulty of a test will be largely influenced by the most difficult question and the last question attempted. If the exam is ordered easy-to-hard, assuming students attempt the questions in order, then the hardest questions will likely be doubly weighted. If the order is reversed, encouraging students to start with the hardest problems and end with the easiest, then the weighting should be shared between the hardest and easiest questions, thus giving the student an overall reduced perception of the difficulty of the test.

In this paper, we first replicate and then extend our previous work which uses students estimation of their own marks on an introductory computer science final examination as a proxy for confidence and difficulty[6]. We support the initial study's hypotheses that there is a strong correlation between predictive ability and exam performance, and that student perception is linked more strongly to previous mark history than to actual exam performance.

SIGCSE '19, February 27-March 2, 2019, Minneapolis, MN, USA

ACM ISBN 978-1-4503-5890-3/19/02…$15.00
https://doi.org/10.1145/3287324.3287398

We fail to replicate the study's finding that male students overestimate marks more than females, however we then extend the study to look at domestic and international students and find that the male overestimation trend is present in domestic students, but not international.

After replicating the initial study, we move on to evaluating the performance and grade estimation of students by question ordering. Our initial finding is that, when taking the class as a whole, there is no evidence of any impact of re-ordering the question. However, when analyzing sub-groups, we find that reversing the traditional ordering of easy-hard may have a negative impact on the performance of international students. Most interestingly, we find that the non-traditional hard-easy ordering has a statistically significant impact on female students, slightly improving their performance, but greatly improving their confidence.

2 BACKGROUND

The effect of question ordering on multiple choice tests has been long studied for both general knowledge[17] and topic specific questions [14]. While results show that the ordering of the questions from the easiest to most difficult results in more optimistic prediction of performance by participant regardless of where in a test participants are asked to estimate their performance, it affects participants' judgment of test fairness [5] and the difficulty rating of the questions [18].

Balch [1] reported that students perform better if questions are given in the same order as the material covered in lectures or presented in textbook as opposed to when the questions are ordered randomly. This result was later contradicted by Neely et al., [12] who ran three experiments with sequential and random order multiple choice question and found no significant differences in terms of performance. They did however find that randomizing the questions increased students' anxiety.

Gender differences in student populations has also been a topic of interest for many research activities ranging from performance evaluation and leadership ability [7, 9] to their attitude towards technology [10, 13]. A few studies have researched the gap between students' predictive and actual performance with regards to their gender[15]. The general findings have been that while there was no significant difference between female and male students' performance in Introduction to Computer Science courses, female students regularly predict their marks significantly lower than male students [6] These results supports the claim that female students underestimate their performance in computer science courses[3]. The reasons behind this low self-efficacy of female students have been discussed in numerous studies. Stereotype threats [4, 16] and lack of female role models [11] are regularly cited as causes.

One part of this study attempts to replicate earlier work[6] where students were asked to predict their own marks after completing the final examination of an introductory computer science course. They found that exam performance was strongly correlated with predictive ability, students with prior programming experience tended to over-predict their marks, and that while there was no difference in mark between male and female students, male students over-predicted significantly more than females.

3 METHODOLOGY

CSCA08: Introduction to Computer Science I is an introductory computer science (CS1) course at The University of Toronto Scarborough focusing on computer science fundamentals for first year students intending to major or specialize in Computer Science. The final exam for the course is a three hour pen and paper exam worth 40% of the final grade, and is cumulative across the entire semester. A total of 243 students completed the course, 62 females and 154 males (the other 27 choosing not to specify). 104 students were domestic, 116 were international, with an additional 23 not permitting access to their status information.

Following the methodology of [6], a bonus question was added to the final exam with the following question:

> Predict your mark on each question of this examination. If your estimate is within 10% of your actual grade, you get this bonus mark.

As with the prior study, marks were awarded to all students who completed the question with sensible results regardless of how close they were to their final marks.

After the non-bonus questions had been prepared by the teaching team[1], the questions were presented to a group of experienced instructors in random order. The instructor were then asked to rank the questions in order of difficulty, as perceived by a CS1 student. The opinions were unanimous with the initial suspicions of the teaching team, and thus a fully ordered ranking of question difficulty was established.

Two exams were prepared, version A (hereafter called easy-difficult) ordered the questions from easiest to most difficult, and version B (hereafter called difficult-easy) ordered the questions from most difficult to easiest. The same questions were used on both tests, and in both tests the bonus questions remained at the end. The only other distinction on the tests was a small **A** or **B** in the corner of the cover page, which, to the best of our knowledge, went entirely unnoticed by the students in the course. Exams were distributed randomly so that each student had an equal chance of receiving either type of exam.

Student gender was determined by the student's choice of honorific, as self declared in the university administration system ('Mr.' or 'M.' for males, 'Mrs.', 'Miss.', 'Ms', 'Mme', or 'Mlle' for female. An additional option of 'Mx.' was offered, but not chosen by any students in this study), students could also choose not to declare an honorific.

Domestic or international status was determined by the student's current tuition status, whether they were paying domestic or international fees for the course.

We did not have access to a student prior programming experience survey, so we were unable to replicate that part of the initial study or compare the effects of question ordering by prior experience.

All data was collected and interpreted post-hoc, and anonymized by members of the study group who had no ability to set or alter student grades. To determine the effect of the test ordering, we performed factorial ANOVA for each of the following: actual grade,

[1] In the interest of full disclosure, two of the members of the teaching team are co-authors on this paper, however they had no role in the marking of the examinations or the statistical calculations of the data

predicted grade, and over-prediction of grade. The ANOVA assumptions hold based on results from *qqnorm* plots, residual plots, Cook's Distance plots, and leverage plots. By Bartlett's test for homogeneity of variance[2], the model has constancy of variance holding under each factor ($p = 0.4835$).

4 RESULTS

In this section we will first explain how we replicated an existing study and then expanded upon it. We will then provide the results for our question ordering study.

4.1 Replication

In the first phase of the study, we attempted to replicate our earlier findings[6] with a new cohort of students. The results will be presented here in brief. For further discussion and analysis, the authors refer the reader to the original paper.

The correlational results were found to hold with similar results. Actual and predicted mark showed a strong correlation (Spearman $\rho = 0.702, p < 0.001$). As did mark with both prediction error and prediction error magnitude ($\rho = -0.292, p < 0.001$ and $\rho = -0.293, p < 0.001$).

As in the initial study, we found that exam under-performers over-predicted their marks more than their counterparts, and the results were strongly statistically significant (over-performer mean = 9.575, under-performer mean = 14.848, t-test $p = 0.0017$).

With respect to gender, we also found no significant difference between the actual grades of male and female students (male mean = 68.301 female mean = 64.486, t-test $p = 0.1379$), but unlike the previous study, we did not find a statistically significant difference in the predictions of male and female students (male mean = 80.308, female mean = 76.955, t-test $p = 0.2112$). While male students did predict slightly higher averages, they also performed slightly better, so that the over-prediction was essentially the same for both groups (male mean = 12.006, female mean = 12.470). However, this may be due to the confounding nature of question ordering and gender which will be further explored in Section 4.3.3.

4.1.1 Extension of Replicated Study. One additional piece of information we included in the study was domestic vs international student status which was not considered in the initial study. We found that in this case, there was no statistically significant difference in either the actual or predicted marks of domestic or international students. It can be discerned from Figure 1 that the predicted and the actual grade of international students in both genders have almost identical means, in particular, the average actual grade for male and female international students are 79.54298 and 79.60192, respectively (t test $p = 0.808$). The average predicted mark for male and female international students are 66.74737 and 65.99038, respectively (t test $p = 0.9864$). Furthermore, the overprediction of grades appeared very similar for both genders. On the contrary, the difference in the average of predicted and actual grade for the domestic students in both genders appeared visibly different from Figure 1. In addition, male domestic students seemed to overpredict their grade compared to the domestic female students.

While we were unable to replicate the effects found in the initial study with respect to gender confidence, separating the data by both gender and student status shows some interesting effects. As seen in Figure 1, the difference in confidence (as estimated by over-prediction of marks) can be clearly seen in domestic students, but was almost non-existent in international students. The average overprediction for male international students is 12.49386, for female international students is 13.05769 (t test $p = 0.824$). While the mean of overprediction for male domestic students is 13.1, and for female domestic students is 10.3825 (t test $p = 0.5926$).

However, none of the differences are statistically significant based on the results reported by a two-sample t-test despite of the obvious differences of the values. Such a result can be attributed to the presence of interaction effects which rendering the t-test an insufficiently robust hypothesis testing procedure.

4.2 Question Ordering

The primary concern of the study was the impact of question ordering on exam performance. To that end, we found that while students did perform slightly better in the easy-difficult ordering group with an average grade of 68.26620 versus an average of 66.21687 for the difficult-easy ordering, the result were not significantly different (t test $p = 0.4192$). Furthermore, we found that the average predicted grades of the two groups were nearly identical with a mean of 79.77817 for easy-difficult ordering group versus a mean of 78.90301 for difficult-easy ordering group. However, the difference is statistically insignificant with a $p = 0.7174$. This runs counter to our initial assumptions, and the generally accepted 'peak-end' rule, which would expect students ending with the easier questions to have perceived the exam as being easier, and thus predict higher marks on average, however as can be seen in Figure 2, the question ordering did not seem to have a large impact in either direction. This may be attributed to the confounding nature of question ordering and gender which demonstrated that the effect of these factors seem to cancel out. The ANOVA results suggested that the interaction effect between gender and question ordering in all students' predicted grade as well as their overprediction of the grade is indeed statistically significant with a p-value of p=0.0135 and p = 0.00415, respectively. If the significance of the interaction effect between gender and question ordering is not neglected, then it must be concluded that the main effect of the corresponding factors, namely, gender and question ordering, on the overprediction of grade is also statistically significant. Furthermore, the interaction variable of gender and question ordering is a potential effect modifier which differentially modifies/influences the observed main effects of these factors on the outcome response variable (whether it is actual grade, predicted grade, or the overprediction of grade) depending on the combinations of factorial levels as seen in Section 4.3.3.

4.3 Sub-Group Effects

Our initial analysis indicates that question ordering has little to no effect on students. However, this is taking the student population as a homogeneous group. Indeed, the main effect of question ordering on students' actual grade is statistically insignificant with the p-value of p=0.342 for the F-test. Unsurprisingly, its F-test has a p-value of 0.342 suggesting statistical insignificance of the main effect of question ordering does not affect students' prediction when considered as a whole without considering sub-groups. Therefore,

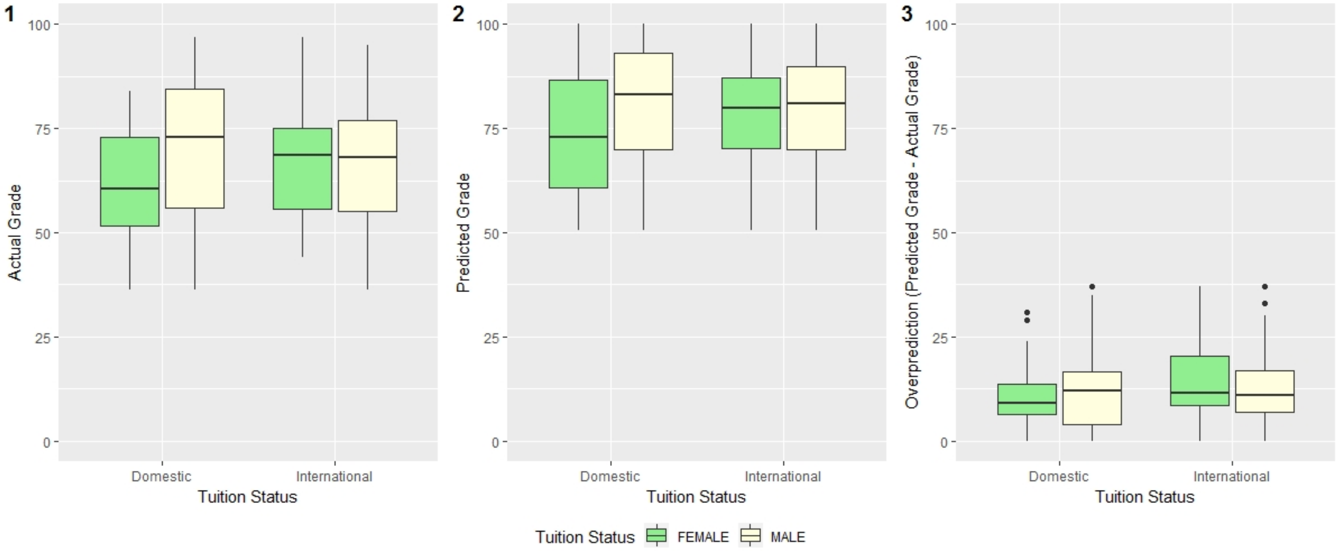

Figure 1: Exam Mark by Tuition Status and Gender

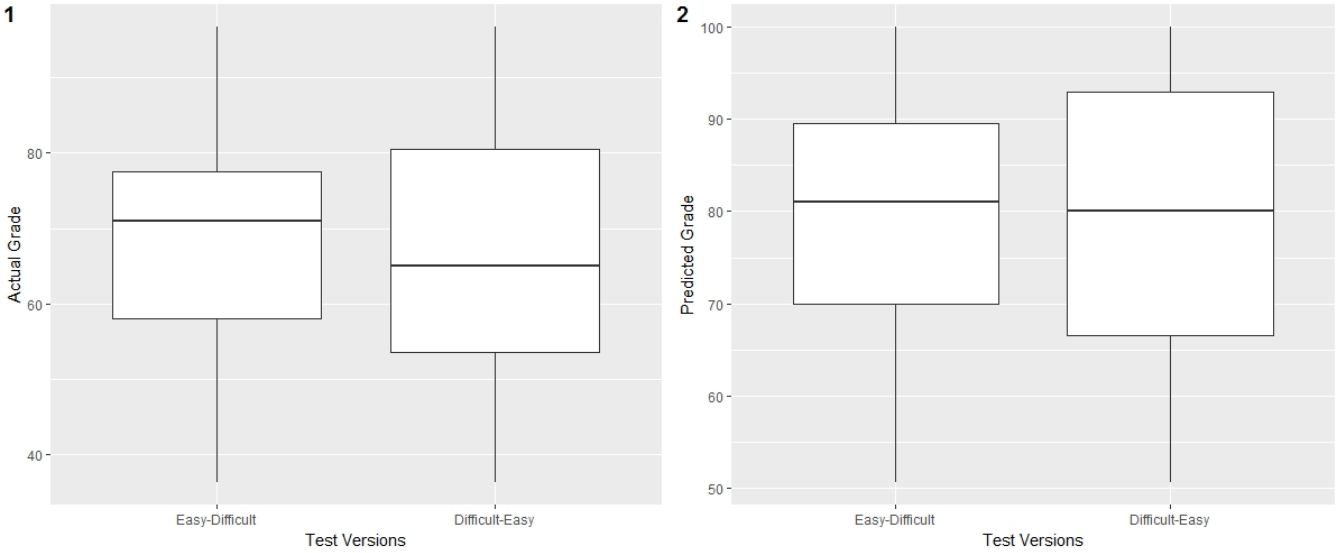

Figure 2: Exam Mark by Question Ordering

we further analyzed the data to determine whether the effect was equally benign in all sub-groups.

4.3.1 Domestic vs International Students. When the data is organized by student status (domestic vs international students) as in Figure 3, we can see that for the traditional easy-difficult ordering, the domestic and international students perform almost identically, but the international students predict slightly higher marks with an average prediction of 80.96047 while the domestic students attained an average of 77.9625. When the question order is reversed, it has very little impact on the domestic students performance, but does appear to boost their confidence slightly with an average prediction of 79.68953 , while it appears to have a negative impact on both the performance and predicted grades for international students who possessed an actual grade of 64.63375 and a predicted grade of 78.0575 on average.

The data in Figure 3 suggest that question ordering appear to affect international students more heavily in comparison to the domestic students. However, we found that in this case, there was no statistically significant difference in the actual marks (F-test p value p=0.493), predicted marks (F-test p value p = 0.6258), and their difference (F-test p-value p = 0.44949) of domestic or international students based on the ANOVA model of regressing each of the

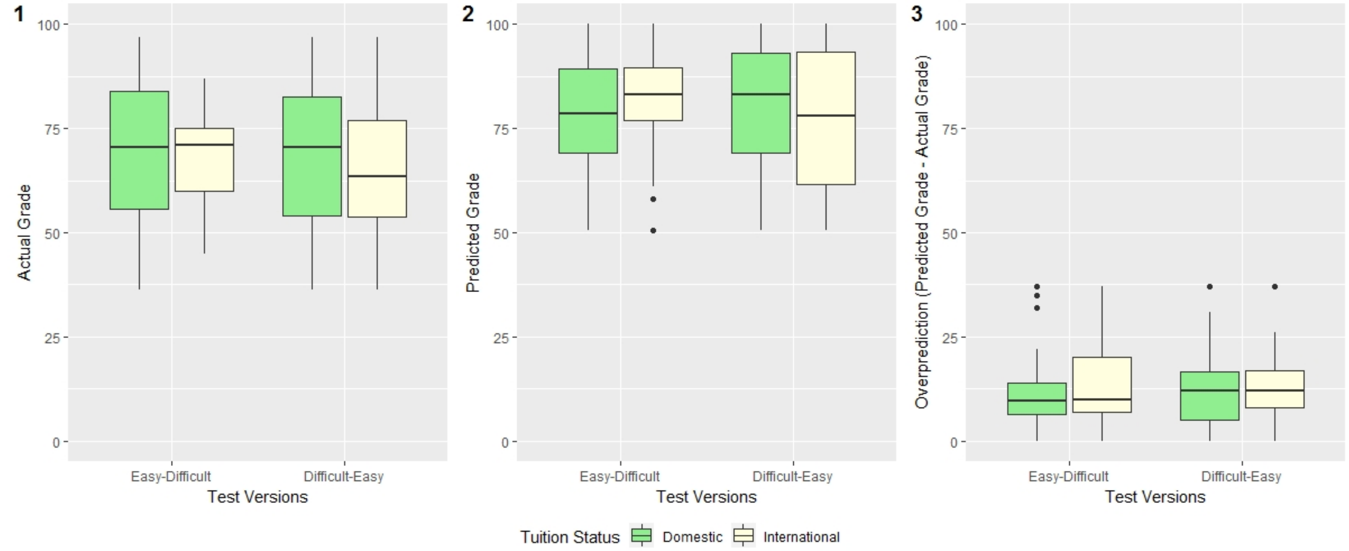

Figure 3: Exam Mark by Question Ordering and Tuition Status

values of predicted mark, actual mark, and their difference on 3 factors: Gender, Tuition Status, and Question Ordering.

4.3.2 Exam Performance Relative to Expectations. As seen in Figure 4, the question ordering had little effect on students when grouped by their performance relative to expectations. It can be discerned that students who have generally performed better throughout the semester continue to perform well on the final exam. Students who underperformed on the final exam seemed to be less affected by the question ordering considering the mean for the students who took the easy-difficult exam obtain a score of 55.410 while predicting 71.73 and the average score of the difficult-easy exam is 54 while predicting 70.52. However, the differences between the average actual score of the two exams with different question ordering for both the underperforming (t test $p = 0.5555$) and the outperforming (t test $p = 0.1831$) individuals are statistically insignificant when stratifying the entire sample by this single factor of performance. Similarly, the difference between the predicted grade of both question ordering groups does not show statistical significance with a t-test $p = 0.7489$ for the underperforming group and $p = 0.1636$ for the overperforming group. This result could be attributed to multicollinearity of the predictors (term test marks) of the regression model that is used to stratify students into outperforming and underperforming groups. Multicollinearity could render the regression model unstable in terms of their prediction power and accuracy. Furthermore, bias can be enlarged and obscure the inferences solely basing on t-tests.

4.3.3 Gender. When the data is organized by gender, the effects of question ordering become quite stark. While the effects on the population overall are negligible, this appears to be because of offsetting opposite impacts on male and female students. While male students appear to have performed better in the more traditional easy-difficult exam setup, reversing the order seems to have all but eliminated the difference in performance the difference of the average of actual grade for female and male in the difficult-easy question ordering has been greatly narrowed. Female students obtained an average of 64.79762 while male students achieved 66.69758 with an average difference of 1.9 on the actual exam grade. Interestingly, while the more traditional ordering clearly showed the strong male over-prediction that we were unable to replicate when analyzing the tests without considering the order, the difficult-easy ordering produced almost identical predicted marks for males and females, and in fact produced the opposite effect, with females over-predicting more than their male counterparts. Since the interaction variable of gender and question ordering is an effect modifier, we cannot interpret the main effects of gender and question ordering separately, rather, it is more justified to conclude that different genders perform very differently when subjected to different test versions. Note that the difference between the predicted grade and actual grade shows a statistically significant interaction term (F-test p-value p = 0.00659) of gender and question ordering, the large difference of grade over-prediction under different test versions for both genders can be attributed to the confounding effects of gender with question ordering and potentially tuition status.

5 CONCLUSIONS

In this paper, we set out to replicate an existing study on confidence and self prediction in CS1 students. We were able to reproduce all of the findings in the areas replicated, aside from the relative over-prediction of marks by male students. However, we then extended this study by subdividing the groups by status (domestic vs international). While this division did not have an impact on its own, the data shows that the gender confidence gap previously found does not hold equally across domestic and international students. However, a larger sample size is required to establish statistical significance.

This study is based on a single examination with a single cohort of students, though it does replicate (most of) the results of our

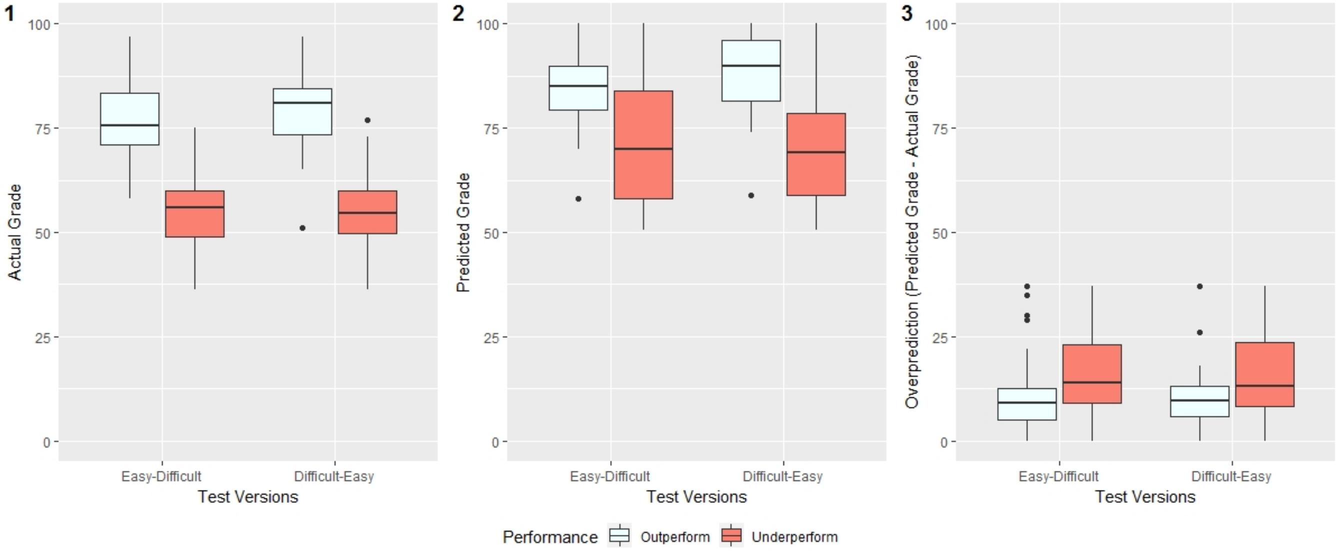

Figure 4: Exam Mark by Question Ordering and Term Test Performance

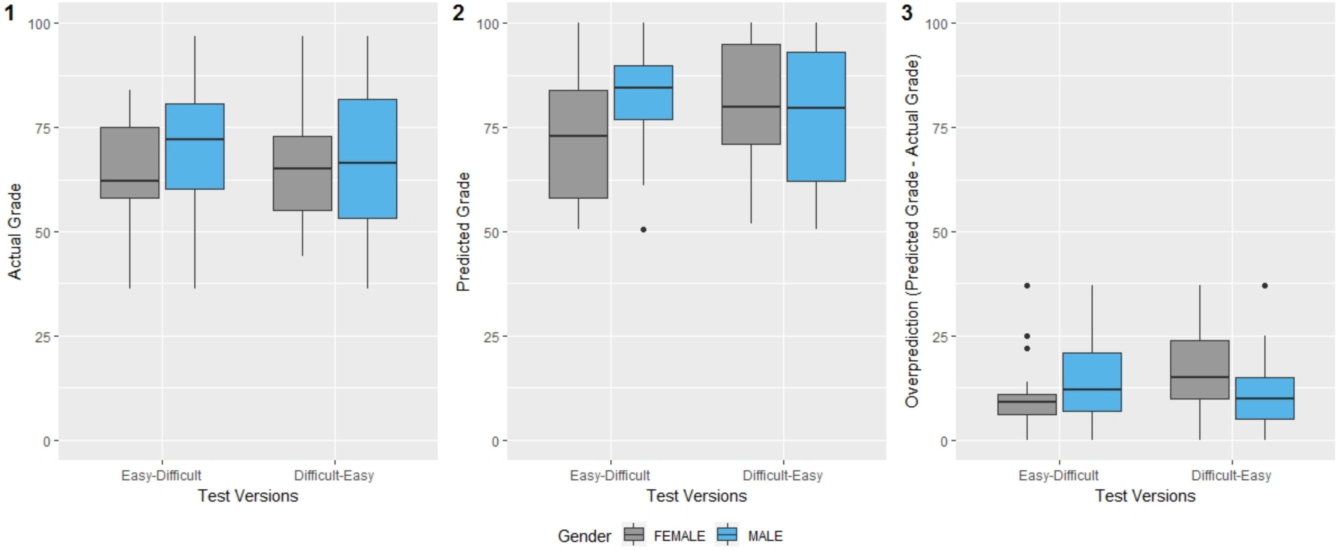

Figure 5: Exam Mark by Question Ordering and Gender

earlier studies. The questions were not specifically designed for the purposes of this study, so it is possible that floor and ceiling effects may have contributed to the results or lack thereof. The data is not conclusive enough to recommend a specific question ordering for all students, nor indeed to say conclusively that there is a global impact of question ordering. The results do show that question ordering can have significant effects on certain sub-groups, but further work and replication will be needed to determine the magnitude and direction of these effects across populations and to begin to provide actionable guidance on question ordering for examinations.

Our initial assumption that we would see an impact of the peak-end rule when re-ordering exam questions proved to be false for the student population as a whole. However, this was because the effect was only present in females, and actually reversed in males. Further work is needed to explain why re-ordering would have an opposite impact on males and females. However it is clear from this initial study that the simple assumption that exams should be structured with the easiest questions first and the hardest questions last needs to be challenged, especially as this ordering seems to have a negative impact on both the performance and confidence of female students.

REFERENCES

[1] William R. Balch. 1989. Item Order Affects Performance on Multiple-Choice Exams. *Teaching of Psychology* 16, 2 (1989), 75–77. https://doi.org/10.1207/s15328023top1602_9 arXiv:https://doi.org/10.1207/s15328023top1602$_9$

[2] Maurice S Bartlett. 1954. A note on the multiplying factors for various χ 2 approximations. *Journal of the Royal Statistical Society. Series B (Methodological)* (1954), 296–298.

[3] Tor Busch. 1995. Gender differences in self-efficacy and attitudes toward computers. *Journal of educational computing research* 12, 2 (1995), 147–158.

[4] Joel Cooper. 2006. The digital divide: The special case of gender. *Journal of Computer Assisted Learning* 22, 5 (2006), 320–334.

[5] Michael L. Dean. 1973. The Impact of Exam Question Order Effects on Student Evaluations. *The Journal of Psychology* 85, 2 (1973), 245–248. https://doi.org/10.1080/00223980.1973.9915653 arXiv:https://doi.org/10.1080/00223980.1973.9915653

[6] Brian Harrington, Shichong Peng, Xiaomeng Jin, and Minhaz Khan. 2018. Gender, Confidence, and Mark Prediction in CS Examinations. In *Proceedings of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education (ITiCSE 2018)*. ACM, New York, NY, USA, 230–235. https://doi.org/10.1145/3197091.3197116

[7] Nagore Iriberri and Pedro Rey-Biel. 2017. Stereotypes are only a threat when beliefs are reinforced: On the sensitivity of gender differences in performance under competition to information provision. *Journal of Economic Behavior Organization* 135 (2017), 99 – 111. https://doi.org/10.1016/j.jebo.2017.01.012

[8] Daniel Kahneman and Richard H Thaler. 2006. Anomalies: Utility maximization and experienced utility. *Journal of Economic Perspectives* 20, 1 (2006), 221–234.

[9] Harsh K. Luthar. [n. d.]. Gender differences in evaluation of performance and leadership ability: Autocratic vs. democratic managers. *Sex Roles* 35, 5 ([n. d.]), 337–361. https://doi.org/10.1007/BF01664773

[10] Anahita Mahzari and Marzieh Ahmadzadeh. 2013. Finding Gender Preferences in E-Commerce Website Design by an Experimental Approach. *International Journal of Applied Information Systems* 5, 2 (2013).

[11] David M Marx and Jasmin S Roman. 2002. Female role models: Protecting womenâĂŹs math test performance. *Personality and Social Psychology Bulletin* 28, 9 (2002), 1183–1193.

[12] Darlene L Neely, Frederick J Springston, and Stephen JH McCann. 1994. Does item order affect performance on multiple-choice exams? *Teaching of Psychology* 21, 1 (1994), 44–45.

[13] Antonio Padilla-MelÃĺndez, Ana Rosa del Aguila-Obra, and Aurora Garrido-Moreno. 2013. Perceived playfulness, gender differences and technology acceptance model in a blended learning scenario. *Computers Education* 63 (2013), 306 – 317. https://doi.org/10.1016/j.compedu.2012.12.014

[14] II Pettijohn, F Terry, and Matthew F Sacco. 2007. Multiple-Choice Exam Question Order Influences on Student Performance, Completion Time, and Perceptions. *Journal of Instructional Psychology* 34, 3 (2007).

[15] Monika Sieverding and Sabine C. Koch. 2009. (Self-)Evaluation of computer competence: How gender matters. *Computers Education* 52, 3 (2009), 696 – 701. https://doi.org/10.1016/j.compedu.2008.11.016

[16] Jessi L Smith and Camille S Johnson. 2006. A stereotype boost or choking under pressure? Positive gender stereotypes and men who are low in domain identification. *Basic and Applied Social Psychology* 28, 1 (2006), 51–63.

[17] Yana Weinstein and Henry L. Roediger. 2010. Retrospective bias in test performance: Providing easy items at the beginning of a test makes students believe they did better on it. *Memory & Cognition* 38, 3 (01 Apr 2010), 366–376. https://doi.org/10.3758/MC.38.3.366

[18] Yana Weinstein and Henry L. Roediger. 2012. The effect of question order on evaluations of test performance: how does the bias evolve? *Memory & Cognition* 40, 5 (01 Jul 2012), 727–735. https://doi.org/10.3758/s13421-012-0187-3