Impact of Online vs. Hybrid Learning Models on Student Performance in Post-Pandemic Statistics Course

Juntao Guo

1-11-2024

24699807

Table of Contents

Abstract

This study assesses the impact of instructional models on student performance, specifically examining the fully online model used during the pandemic and the post-pandemic hybrid model that combines online teaching with on-campus exams. Using recent course data, this research also evaluates whether allowing two test attempts affects student outcomes and explores the data's adequacy for predictive modeling of final exam scores.

Key findings show that both teaching models significantly affect overall performance, notably enhancing outcomes among high-achieving students (HD group). A strong association is observed between third and fourth test scores and final exam results, and the option for dual test attempts positively influences performance. Initial modeling efforts achieved an accuracy of approximately 40%, indicating that future predictive reliability would benefit from additional data features.

1. Introduction

Since the onset of the pandemic, universities worldwide transitioned to fully online teaching as campuses closed and in-person learning was restricted. This shift significantly altered how students engaged with their coursework, leading to numerous studies comparing online learning with traditional in-person instruction. Many of these studies reported that students performed better in fully online settings.

As the pandemic subsided, universities reopened, adopting a hybrid model that combined online learning with in-person instruction and assessments. This hybrid model differs from both pre-pandemic fully in-person formats and pandemic-era online models, potentially impacting student performance in new ways. With greater access to face-to-face interactions, students may find it easier to ask questions and participate in discussions. Additionally, the shift back to in-person assessments may influence test outcomes.

This study focuses on a statistics course that integrates statistical theory and foundational mathematics. Course assessment comprises four Mystery Tests and a final exam, with overall grades calculated as 0.05M1 + 0.15M2 + 0.15M3 + 0.15M4 + 0.5*Final. Students are allowed two attempts per Mystery Test, with the highest score recorded as the final result. The first test assesses prerequisite knowledge beyond the semester's material, while the remaining tests progressively evaluate material covered in Weeks 1–3, Weeks 4–6, and Weeks 7–10, addressing both foundational math and specialized statistical concepts.

The data for this study comes from the same course, taught by the same professor, during two semesters: 2021 (fully online) and 2023 (post-pandemic hybrid). The 2023 dataset includes both test attempts for each student, allowing for analysis of the effect of multiple test attempts on performance. In line with UTS grading guidelines, student grades are classified into five categories—High Distinction (HD), Distinction (D), Credit (C), Pass (P), and Fail (Z)—based on score ranges of 85–100%, 75–84%, 65–74%, 50–64%, and 0–49%, respectively. This classification provides insights into performance changes across different grade levels under each instructional model.

The study aims to address the following research questions:

1. Does the transition from fully online teaching to the hybrid model affect students' overall performance?

2. Has the shift in teaching methods affected students' test performance during the semester, and how are the Mystery Tests related to final exam outcomes?
3. In the new hybrid teaching model, is offering two test attempts important or impactful?
4. Can a predictive model be developed based on the new teaching model to forecast final exam scores using test results for future semesters?

2. Material and methods

2.1 Data Collection

This research dataset was collected by the University of Technology Sydney through its internal system, with the data stored in Excel format. The participants in this study were students enrolled in the course during two semesters: 2021, which was taught during the pandemic, and 2023, which was taught using a hybrid model of online and in-person instruction. The sample includes exam information from 676 students from 2021 and 874 students from 2023. Both samples contain data on student performance in the tests, the final exam, and the overall course grades. Notably, the 2023 dataset provides more detailed information, including scores from both attempts for each test, while the 2021 dataset only records the highest score from the two attempts.

2.2 Data Pre-processing

The data pre-processing step included data quality assessment, data cleansing, data transformation, and data reduction. During this process, missing values (originally stored as "NaN") in the student scores were replaced with zeros. Additionally, students whose scores were consistently zero across all assessments were removed from the dataset, as this group could not provide useful insights for the study and were considered outliers. The removal of these outliers enhanced the accuracy of subsequent data analysis and improved the performance of the models developed later. After this step, 667 students from 2021 and 861 students from 2023 were retained in the dataset.

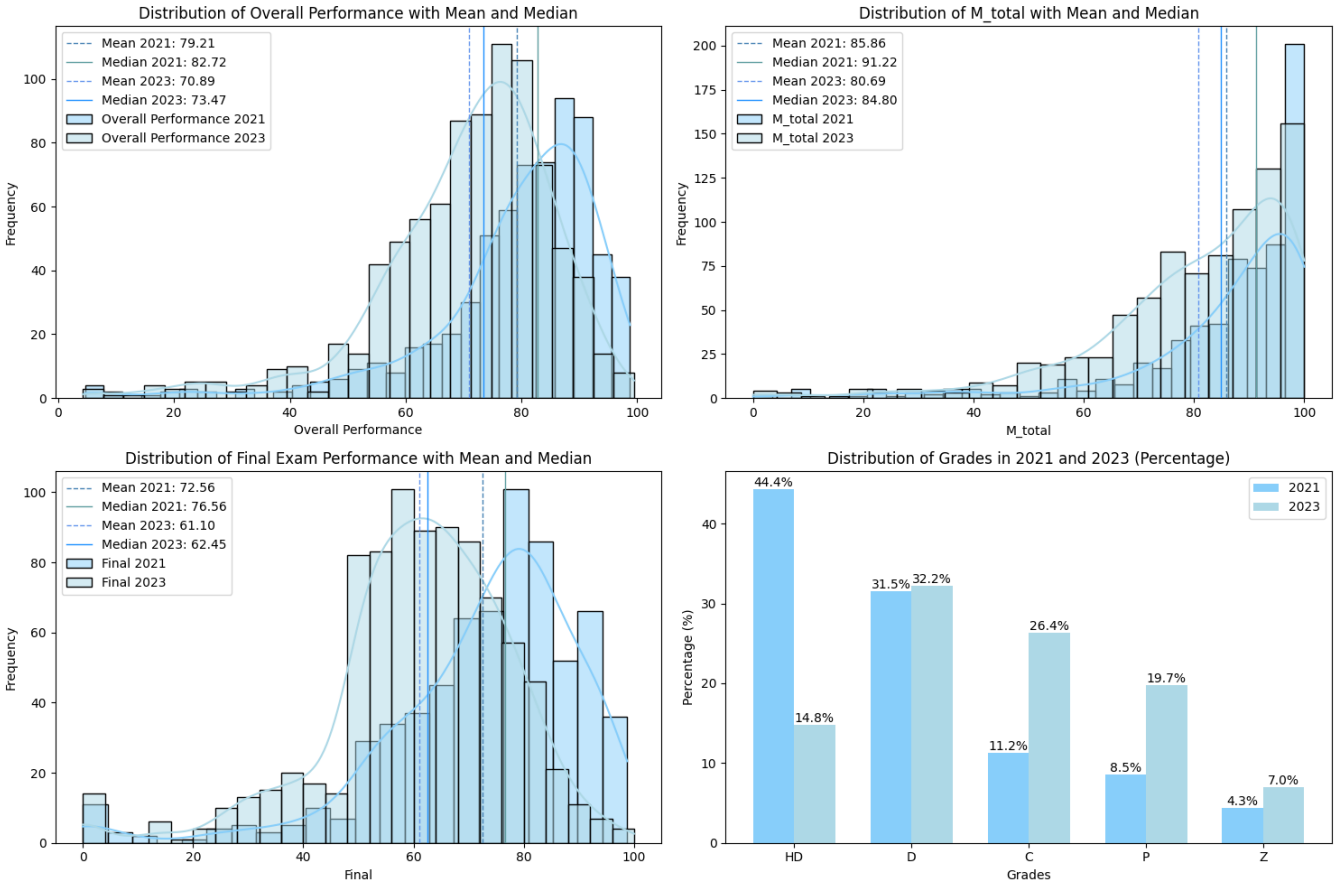
Furthermore, based on the final weightings of the test scores, a new metric column named "M_Total" was created to represent the overall test performance of each student. This was calculated using the formula: 0.05M1 + 0.15M2 + 0.15M3 + 0.15M4. Additionally, based on the university's grading system, each student's overall grade was categorized into five groups: High Distinction (HD), Distinction (D), Credit (C), Pass (P), and Fail (Z), corresponding to score ranges of 85-100%, 75-84%, 65-74%, 50-64%, and 0-49%, respectively. For the 2023 dataset, an additional column was introduced for each test, classifying students based on their participation in the two test attempts: (1) only attempted the first test, (2) only attempted the second test, or (3) attempted both tests. This classification will be used to assess the importance and impact of offering two test opportunities and further construction of the prediction model.

2.3 Data Analysis – Distribution

The distributions of student performance for both 2021 and 2023 are presented in the histograms in **Fig. 1**, which depict the overall performance, average mystery test score, and final exam scores. The histograms clearly show that all distributions

exhibits left-tailed skewness, indicating that a larger number of students scored toward the lower end of the spectrum. Moreover, notable differences can be observed in the distributions, means, and medians between 2021 and 2023. These differences suggest that the transition from the fully online teaching model during the pandemic to the hybrid model after the pandemic may have contributed to a decline in student performance.

Figure. 1 Distribution of Overall Performance, Total Mystery Test Score, Final and Grade



Based on the characteristics observed in **Fig. 1**, it is evident that the data follows a non-normal distribution. To rigorously confirm whether the differences in the distributions of the two datasets (2021 and 2023) are statistically significant, this research selected two non-parametric tests, the Mann-Whitney U test and the Kolmogorov-Smirnov test. These non-parametric tests are ideal for comparing distributions when data does not follow a normal distribution, as they do not rely on the assumption of normality, making them robust for skewed or irregularly distributed data.

These tests were applied by pairing the data from 2021 and 2023 for overall performance scores. As reported in Table 1, the results from both tests indicate significant differences at the $p < 0.05$ level for all three metrics. This finding suggests that the shift from a fully online teaching mode to the hybrid teaching and assessment mode significantly impacted student test scores and final exam performance, resulting in a decline in their overall performance.

Table 1. The difference of overall performance for 2021 and 2023

	2021	2023
Mean	79.21	70.89
Median	82.72	73.47
Mann-Whitney U Test	Statics: 402515.5, P-value: 1.86*e-41	
Kolmogorov-Smirnov test	Statics: 0.3362, P-value: 3.57*e-38	
Chi-Square Test	Statics: 198.86, P-value: 6.59*e-42	

Since the results indicate a significant decline in student performance between 2021 and 2023, additional analysis was conducted to explore how different grade categories were affected by the transition in teaching and assessment methods. In this study, student performance was classified into five grade groups: High Distinction (HD), Distinction (D), Credit (C), Pass (P), and Fail (Z), based on the university's grading system. **Fig. 1** illustrates the distribution of these grades for both 2021 and 2023. Because the number of students differs between the two years, the scores in each grade category also were converted into proportions to conduct further Chi-Square Test. Notably, in **Fig.1**, except for the High Distinction (HD), all other grade categories exhibited an increase between the two years, reflecting a shift in the distribution of student performance.

To further investigate the observed changes, a Chi-Square test was conducted for grade category to test whether the distributions of grades were significantly different between 2021 and 2023. The following hypotheses were formulated:

- **Null Hypothesis (H_0)**: The distribution of grades, is independent of the year (i.e., there is no significant difference in the grades between 2021 and 2023).
- **Alternative Hypothesis (H_1)**: The distribution of grades is not independent of the year (i.e., there is a significant difference in the grades between 2021 and 2023).

As shown in **Table 1,** the P-value < 0.05 supports the rejection of the null hypothesis, confirming that there are significant differences in the distribution of grades between 2021 and 2023. Moreover, using the standardized residuals, whether there is a significant change between the two years in different grade categories can be determined by checking if the residual values are greater than or less than 2.0.

Specifically, as shown in **Table 2**. The HD, C and P categories show significant differences, with a clear decrease in the proportion of students achieving higher grades (HD), and an increase in the number of students falling into lower grade categories (C, P). The grade category (D and Z) show there is not a statistically significant difference between 2021 and 2023. This suggests that the transition from fully online to hybrid teaching and assessment methods may have disproportionately affected students across different performance levels, leading to a notable decline in high-achieving students and an increase in lower-performing students.

Table. 2 Standardized Residuals for different overall performance grade categories

	2021	2023
HD	8.1946	-7.2126
D	-0.1773	0.1560
C	-4.9495	4.3564
P	-4.2283	3.7215
Z	-1.5803	1.3909

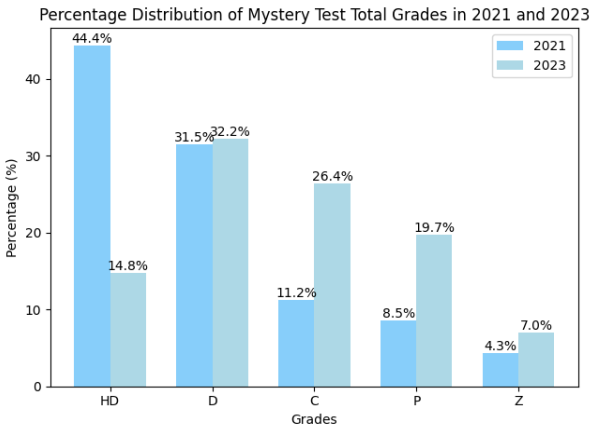
Therefore, this represents that the shift in teaching methods and assessment structures has had a substantial impact on student outcomes, as evidenced by the significant differences in overall performance and the distribution of grades. These findings underscore the importance of understanding how changes in instructional modes can affect student learning and assessment, and they offer valuable insights for future curriculum and teaching model adjustments.

Furthermore, with the transition in teaching methods, all four semester tests shifted from being conducted entirely online to offline. The testing environment in online tests often differs significantly from that of offline tests, which may affect students' performance. This study will examine the total test scores from both years to investigate the potential changes brought about by this shift, particularly to see if the transition affects students differently across various grade groups.

Table 3. Standardized Residuals for different total test grade categories

	2021	2023
Chi-Square Test	Statics: 57.84, P-value: 8.25e-12	
	Standardized Residuals (2023 compared to 2021)	
HD	3.4829	-3.0655
D	-2.1505	1.8928
C	-3.0413	2.6768
P	-2.5204	2.2183
Z	-0.4831	0.4252

Figure 2.Percentage of different Total Test Grade Category



As shown in **Fig. 2**, the proportion of students in each grade category for average mystery tests is compared between 2021 and 2023. Clear differences can be observed in the distribution of students across different score ranges. A chi-square test was also conducted to determine whether the changes in the grade distributions between the two years are statistically significant. The results, presented in **Table 3**, show that there are significant differences in the HD, C, and P categories. However, no significant difference was found in the D and Z grade categories.

These findings suggest that the shift from online to offline exam environments has a notable impact on students' performance in the HD, C, and P score ranges. According to the standardized residuals, the transition from online to offline exams resulted in a significant decrease in the number of high-achieving students (HD). Conversely, there was a significant increase in

students who have a moderate understanding of the material but are not yet proficient (C and P categories). However, for students who have a relatively proficient understanding of the material (D category) and did not engage with the course material or failed to grasp the content (Z category), the change in test format from online to offline did not result in a significant difference.

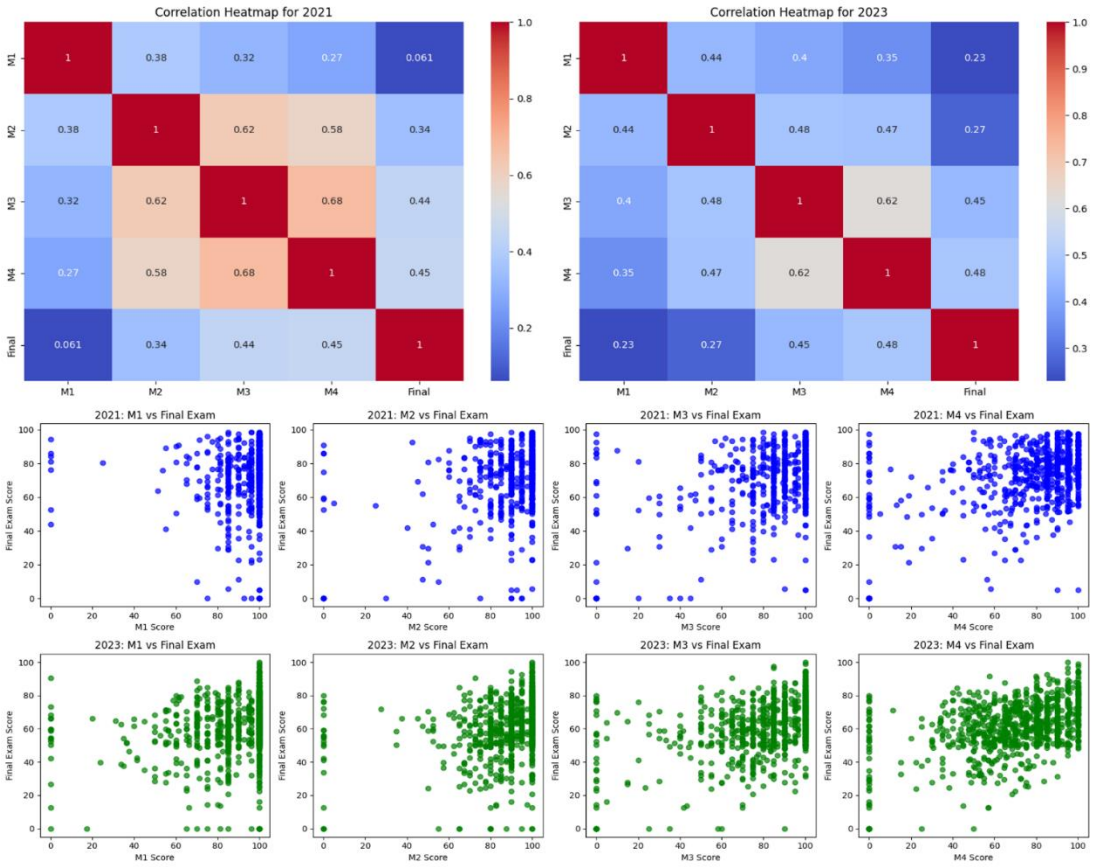
2.4 Data Analysis – Correlation between Mystery Test and Final Exam

Through the chi-square analysis, this study found that changes in the test environment had a significant impact on student performance, identifying which grade groups were positively or negatively affected. This outcome reflects, to some extent, that students' overall understanding of the course material was influenced by the change in teaching methods.

Additionally, the study examined the relationship between each test (M1, M2, M3, M4) and the final exam under different teaching and test environments, exploring how this relationship evolved with the shift in teaching methods.

Correlation analysis and scatter plots were used to investigate the connection between each test and the final exam. In the correlation heatmap shown in **Fig. 3**, it is evident that in both 2021 and 2023, M3 and M4 scores are more strongly correlated with final exam results, showing relatively higher correlations. In contrast, M1 and M2 displayed lower correlations with the final exam. This can be explained by the course structure: M1 is designed to assess prerequisite knowledge for the course, which is why its correlation with the final exam was only 0.061 in 2021 and 0.23 in 2023, the lowest correlation among all the tests.

Fig. 3 Correlation and Scatter Plot between Each Tests and Final Exam



Moreover, the correlation heatmap reveals that each test has the highest correlation with the preceding test, indicating the progressive nature of the course content. As the course advances, the material builds on previous knowledge, and strong performance on one test suggests a higher likelihood of success on subsequent tests. This pattern is further supported by the scatter plots in **Figure. 3**, where an upward trend from M1 to M4 is clearly visible across both years, indicating a stronger predictive relationship. This suggests that performance on later tests (M3 and M4) is more predictive of final exam success.

2.5 Data Analysis – The Impact of Two Test Opportunities on Student Performance

In this study, each Mystery test offered students two opportunities to participate. As shown in the histogram in **Figure 4**, there are notable differences in the distribution of Mystery test scores depending on whether students participated only in the first test, only in the second test, or in both tests. The Shapiro-Wilk test for Mystery test scores (p-value < 0.05) indicated that the 2023 Mystery test scores followed a non-normal distribution. Additionally, a significant correlation was observed between the M3 and M4 test scores and the final exam scores. Consequently, we used the Kruskal-Wallis H test to examine whether the number of test attempts significantly affected test and final exam scores for M3 and M4.

Figure 4. Scatter Plot, Line plot with error bars and Histogram for test attempt in Test 3 and Test 4

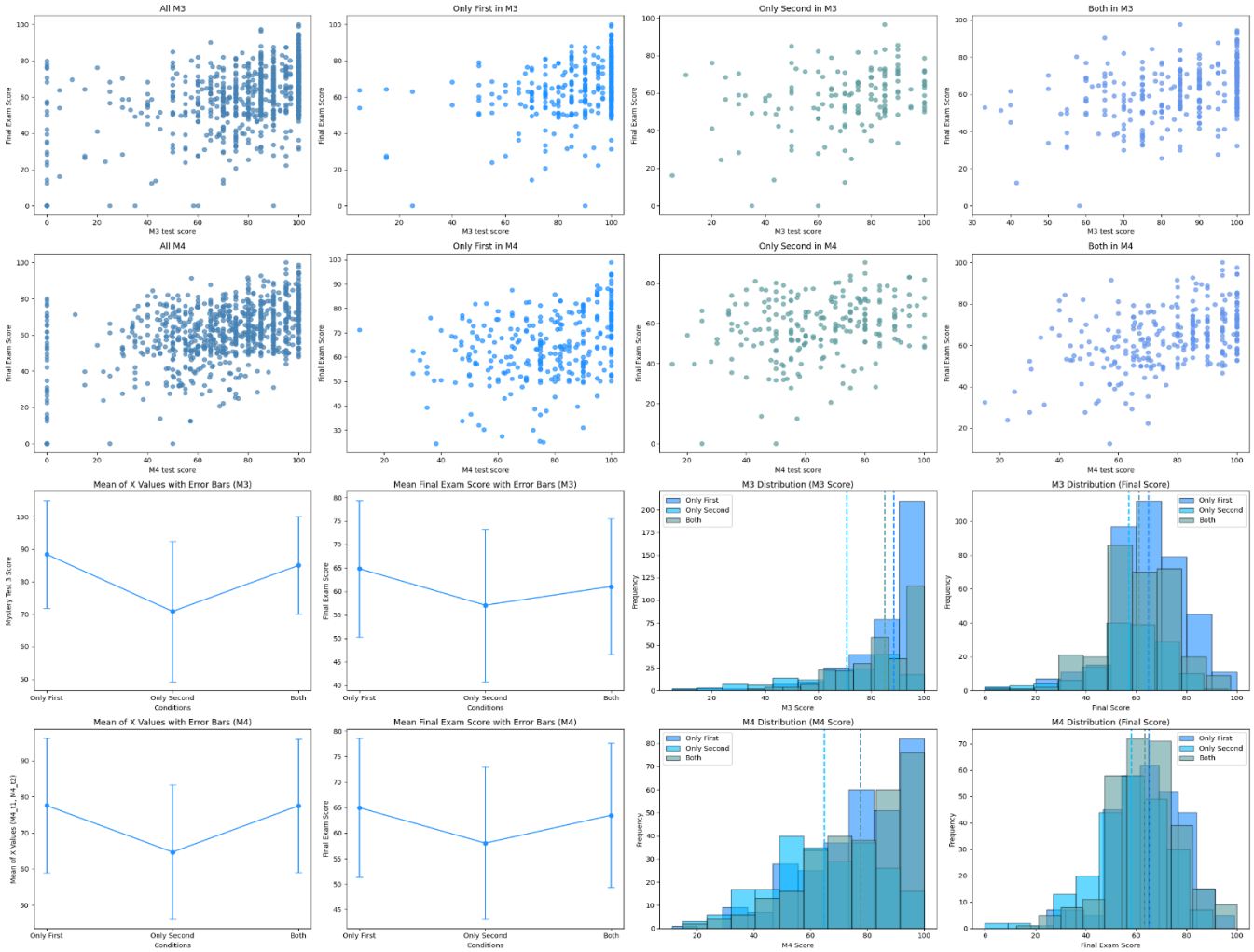


Figure. 4 includes scatter plots showing that students who only participated in the first test generally performed better than those who only participated in the second test, achieving higher scores in both the test and the corresponding final exam. This suggests that taking tests in alignment with the course schedule may enhance final exam performance. In contrast, students in the 'Only Second' group displayed greater variability in both test and final exam scores, which may reflect inconsistent preparation or performance levels.

A positive correlation between test scores and final exam results across all groups reinforces the idea that higher test scores are generally predictive of better final exam outcomes. However, the presence of outliers, where some students performed well on tests but scored low on the final exam, suggests that additional factors, such as exam conditions or preparation strategies, may also influence final exam performance.

Furthermore, as illustrated in **Figure 4**, the line plot with error bars and the histogram showing the distribution of test and final exam scores reveal that students who participated in either only the first test or both tests achieved higher scores compared to those who only took the second test. This trend is consistent across M3, M4, and the final exam. The Kruskal-Wallis H test results presented in **Table 4** confirm that these differences are statistically significant. Post-hoc tests were then conducted to identify specific group pairs that differed. For both M3 and M4, post-hoc pairwise comparisons show p-values < 0.05 for scores of students who took either only the first test or both tests compared to those who only took the second test, indicating a significant difference. This suggests that offering two test opportunities has a positive effect on test performance. However, for final exam scores, participation in M3, whether by taking both tests and only the second test, did not show a significant effect, as the post-hoc pairwise comparison yielded a p-value of 0.074 (> 0.05).

Table 4. Kruskal-Wallis H Test and Post-hoc Pairwise Comparisons for test attempt in Test 3 and Test

M3	**Kruskal-Wallis H Test for Test Score: Statistic: 102.08, P-value: 6.80e-23**			**Kruskal-Wallis H Test for Final Exam Score: Statistic: 30.66, P-value: 2.20e-07**		
Post-hoc Pairwise Comparisons	**Only First**	**Only Second**	**Both**	**Only First**	**Only Second**	**Both**
Only First	1	1.7e-23	0.0004	1	6e-7	0.00091
Only Second	1.7e-23	1	2.8e-11	6e-7	1	0.074
Both	0.0004	2.8e-11	1	0.00091	0.074	1
M4	**Kruskal-Wallis H Test for Test Score: Statistic: 74.12, P-value: 8.05e-17**			**Kruskal-Wallis H Test for Final Exam Score: Statistic: 26.61, P-value: 1.67-06**		
Post-hoc Pairwise Comparisons	**Only First**	**Only Second**	**Both**	**Only First**	**Only Second**	**Both**
Only First	1	4.5e-14	1	1	1e-06	0.59
Only Second	4.5e-14	1	1.4e-13	1e-06	1	0.00046
Both	1	1.4e-13	1	0.59	0.00046	1
Mann-Whitney U Test for Student Taking Test Two Times (First vs Second)						
M3	Statistic: 20813.0, P-value: 3.64e-31					
M4	Statistic: 25887.5, P-value: 2.55e-15					

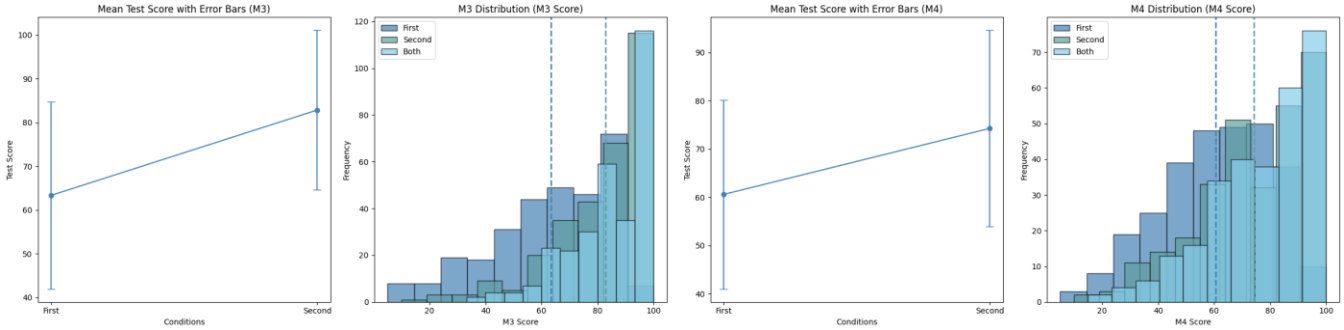
These findings suggest that students who took the first test, either as their only attempt or alongside the second test, performed better, likely because the test was administered shortly after the material was taught, when familiarity with the content was highest. In contrast, students who only took the second test, conducted a week later, may have found it more

challenging due to potential declines in content retention and new material introduction, which may have limited their preparation time.

Although offering two test opportunities significantly impacts test scores, it does not definitively influence final exam performance. While both the Kruskal-Wallis test and post-hoc pairwise comparisons for M4 show p-values < 0.05, indicating a significant difference, no significant difference was found for M3 when comparing final exam scores between students who took both tests and those who only took the second test. Therefore, it cannot be concluded that two test attempts have a significant impact on final exam performance, though they do have a significant impact on test scores.

Interestingly, for students who took both tests, **Figure. 5** shows that they tended to achieve higher scores on the second attempt, with their overall score benefiting from the higher score on the second test. This indicates that students who were more engaged in the course (i.e., those who participated in both tests) benefited from having two assessment opportunities, which allowed them to improve their performance.

Figure. 5 Line plot with error bars and Histogram for student who attend the test twice in Test3 and Test4



Therefore, students who only participated in the second test performed lower not only in that test itself compared to those who took the first test or both tests but also in the final exam. The Mann-Whitney U test results further validated that, for both M3 and M4, students who only took the second test scored significantly lower than those who took the first test or both tests.

In summary, these findings highlight the benefits of offering two test opportunities. Students who struggled in the first test were given the chance to improve in the second attempt, demonstrating the positive impact of repeated assessment. Additionally, the results underscore the importance of maintaining engagement with the course content. Students who took the first test (whether they took only the first test or both tests) consistently outperformed those who only took the second test.

3. Construction of Different Prediction Model

Based on the findings above, this study demonstrates that transitioning from online to offline teaching modes has a measurable impact on student performance. Given that hybrid teaching models are likely to become common in the future, the final objective of this study is to use data from the 2023 hybrid model to develop a predictive model for forecasting student final exam scores under hybrid instructional conditions.

The model incorporates the original dataset features, which include scores from M1, M2, M3, M4, and the final exam, along with four newly added features representing the number of test attempts made by each student. These values are coded as follows: 1 for students who took only the first attempt, 2 for those who took only the second, and 3 for those who attempted both. The dependent variable is the students' final exam grade, categorized into performance levels (HD, D, C, P, Z) based on their final scores.

To develop and validate the predictive model, 80% of the 2023 data was randomly selected as the training set, with the remaining 20% reserved as the test set to assess model accuracy. Table **5** summarizes the performance results of each model tested.

Table 5. Predictive Model and its Accuracy

Model	Accuracy
Logistic Regression	0.45
Random Forest	0.37
XGBoost	0.37
Decision Tree	0.27
Bagging	0.35
AdaBoost	0.27
Gaussian Naive Bayes	0.33
Neural network (RMSProp optimizer, 4 neurons, stack size: 64, 500 epochs)	0.45

The model accuracies were relatively low, with both the neural network and logistic regression achieving the highest accuracy at only 45%. This level of accuracy is insufficient for reliable future predictions. Therefore, additional data features should be considered in future iterations to improve model performance, such as each student's attendance record, assignment scores, and other relevant academic metrics. Expanding the feature set will better enable the model to capture the critical factors that influence performance, resulting in a more accurate and dependable predictive tool.

4. Conclusion and Future works

This study underscores the substantial impact of transitioning from a fully online teaching model during the pandemic to a hybrid model in a post-pandemic environment on student performance in a statistics course. Notably, this shift has influenced overall grades and the distribution across performance categories, suggesting that teaching models play a crucial role in student outcomes. The hybrid model's dual-test attempt policy particularly contributed to performance gains, allowing students additional opportunities to solidify their understanding, which correlated with improved scores on later tests and the final exam. Moreover, the study identified distinct patterns among performance levels: high-achieving students (HD) showed a decline under the hybrid model, while moderate achievers (C and P) experienced gains. This suggests that the assessment format's impact may vary depending on students' proficiency levels, highlighting the need for adaptable strategies in instructional design.

For future research, expanding the dataset with additional variables, such as engagement metrics, attendance, and assignment scores, could improve model accuracy and reliability for predictive analysis. Additionally, exploring longitudinal data across

multiple semesters would provide insight into the lasting impact of hybrid instructional formats. These steps will enable a more robust understanding of instructional methods, ultimately informing the design of teaching models that enhance learning outcomes for diverse student needs and optimize academic success across evolving educational landscapes.