Jurnal Pendidikan Matematika Indonesia
Volum 11 Nomor 1 bulan Maret 2026 ; Page 95 - 108
p-ISSN: 2477-5967 e-ISSN: 2477-8443

Eksplorasi Dinamika Prestasi Akademik pada Evaluasi Akhir Semester Berdasarkan Kelas dan Gender Menggunakan Pendekatan ANOVA Dua Arah

Exploring the Dynamics of Academic Achievement in End of Semester Assessment Based on Class and Gender Using a Two Way ANOVA Approach

Wigbertus Ngabu[1*], Putri Yuanita[2], Elmawati[3], Suci Andriani[4], Yusa Putra[5]

[1,2,3,4,5]Mathematics Education Study Program, Faculty of Teacher Training and Education, Universitas Riau

**Corresponding author: Pekanbaru, Riau, 28293, Indonesia)*
wigbertus.ngabu@lecturer.unri.ac.id[1*]
putri.yuanita@lecturer.unri.ac.id[2]
Elmawati@lecturer.unri.ac.id[3]
suciandriani@lecturer.unri.ac.id[4]
yusaputra@lecturer.unri.ac.id[5]

Received 8 January 2026; Received in revised form 7 February 2026; Accepted 12 February 2026

Kata Kunci :
Prestasi akademik mahasiswa; Statistika dasar; Evaluasi pembelajaran; ANOVA

ABSTRAK

Evaluasi hasil belajar merupakan komponen penting dalam pendidikan tinggi karena mencerminkan capaian pembelajaran serta efektivitas proses pembelajaran. Penelitian ini bertujuan menganalisis dinamika prestasi akademik mahasiswa pada ujian akhir semester mata kuliah Statistika Dasar berdasarkan faktor kelas (A, B, C) dan gender (laki-laki, perempuan). Penelitian menggunakan pendekatan kuantitatif eksplanatori dengan desain faktorial dan dianalisis melalui Analisis Varians Dua Arah (two-way ANOVA). Uji asumsi menunjukkan residual berdistribusi normal (Shapiro–Wilk $p = 0.0854$) dan varians antar-kelompok homogen (Levene p-value = 0.6468), sehingga prasyarat ANOVA terpenuhi. Hasil ANOVA menunjukkan bahwa faktor kelas berpengaruh signifikan terhadap skor ujian akhir semester ($F(2,104) = 2.956$; $p = 0.00562$) dan faktor gender juga berpengaruh signifikan ($F(1,104) = 0.704$; $p = 0.0032$), sedangkan interaksi kelas×gender tidak menunjukkan pengaruh yang signifikan ($F(2,104) = 1.355$; $p = 0.0626$). Uji lanjut Tukey HSD mengungkapkan bahwa perbedaan capaian antar kelas terutama didorong oleh perbedaan Kelas C dan Kelas A (p-adj = 0.0443), sementara perbedaan berdasarkan gender menunjukkan kecenderungan yang konsisten di seluruh kelas ($p - adj = 0.0069$). Temuan ini diperkuat oleh boxplot dan interaction plot yang menampilkan pola distribusi skor sejalan dengan hasil inferensial. Secara keseluruhan, penelitian ini menunjukkan bahwa variasi prestasi akademik mahasiswa dalam evaluasi akhir semester dipengaruhi secara independen oleh faktor kelas dan gender. Hasil penelitian ini diharapkan

dapat menjadi dasar pengambilan keputusan akademik yang lebih berbasis data, serta memberikan kontribusi metodologis dalam penerapan analisis statistik. Berdasarkan temuan tersebut, program studi disarankan melakukan standardisasi implementasi pembelajaran dan evaluasi antar kelas serta menyediakan dukungan belajar yang lebih terarah bagi kelompok mahasiswa yang konsisten menunjukkan capaian lebih rendah agar kesenjangan hasil belajar dapat ditekan secara berbasis data.

Keywords *:*
Student academic achievement; Basic statistics; Learning evaluation; ANOVA

ABSTRACT
The evaluation of learning outcomes is an essential component of higher education because it reflects learning achievement and the effectiveness of the instructional process. This study aims to analyze the dynamics of students' academic performance in the final semester examination of the Basic Statistics course based on class (A, B, C) and gender (male, female). The study employed an explanatory quantitative approach with a factorial design and was analyzed using a two-way analysis of variance (two-way ANOVA). Assumption testing indicated that the residuals were normally distributed (Shapiro–Wilk $p = 0.0854$) and that variances across groups were homogeneous (Levene p-value = 0.6468), confirming that the ANOVA prerequisites were satisfied. The ANOVA results showed that class had a statistically significant effect on final examination scores ($F(2,104) = 2.956$; $p = 0.00562$) and that gender also had a statistically significant effect ($F(1,104) = 0.704$; $p = 0.0032$), whereas the class × gender interaction was not statistically significant ($F(2,104) = 1.355$; $p = 0.0626$). Post hoc analysis using Tukey's HSD revealed that differences across classes were primarily driven by the contrast between Class C and Class A (p-adj = 0.0443), while gender-based differences showed a consistent pattern across classes (p-adj = 0.0069). These findings were supported by boxplots and interaction plots that displayed score distributions consistent with the inferential results. Overall, the study indicates that variation in students' academic performance in the final semester evaluation is independently influenced by class and gender. The findings are expected to inform more data-driven academic decision-making and contribute methodologically to the application of statistical analysis. Based on these results, the study program is advised to standardize instructional implementation and evaluation across classes and to provide more targeted learning support for groups of students who consistently demonstrate lower achievement, so that learning gaps can be reduced in a data-driven manner.

INTRODUCTION

The evaluation of learning outcomes constitutes an essential component of the higher education system, as it serves as an indicator of learning achievement, instructional effectiveness, and the quality of the curriculum implemented (Hamilton et al., 2021). In the context of mathematics education, academic evaluation not only assesses conceptual mastery but also reflects students' logical, analytical, and quantitative thinking abilities (Rutenberg et al., 2022) (Al Hazaa et al., 2021). Consequently, the analysis of students' academic achievement represents a critical aspect in efforts to enhance instructional quality and strengthen institutional accountability.

Final Semester Examinations play a strategic role as summative assessment instruments designed to comprehensively measure students' competency attainment after the completion of the instructional process (Ali, 2024). Final Semester Examinations scores are frequently used as a basis for academic decision-making, both at the individual student level and at the program level (Ismail et al., 2022). Nevertheless, final examination results often exhibit substantial variability across student groups, indicating the presence of academic performance dynamics that warrant more in-depth investigation (Uzun et al., 2025).

In mathematics education programs, particularly in the Basic Statistics course, instructional challenges become increasingly complex. This course requires not only conceptual understanding but also numerical proficiency and strong data interpretation skills. Moreover, Basic Statistics often

functions as a gateway course: difficulties in probability reasoning, statistical inference concepts, and the interpretation of variability and uncertainty can accumulate, potentially affecting students' readiness and persistence in subsequent statistics-related courses and research-methods modules in later semesters. Differences in students' academic backgrounds, learning strategies, and classroom learning dynamics have the potential to influence learning outcomes. Therefore, evaluating academic achievement in the Basic Statistics course necessitates an analytical approach capable of capturing variations and patterns of differences in an objective and systematic manner.

Previous studies have predominantly evaluated student learning outcomes using descriptive approaches or single-factor analyses, which tend to provide only a partial representation of inherently multidimensional academic phenomena (Rabattu et al., 2023). Such approaches are insufficient to fully explain the interactions among factors that simultaneously influence students' academic achievement (Yu et al., 2022). Within the context of learning evaluation, this limitation may hinder a comprehensive understanding of the actual dynamics underlying observed learning outcomes (Guo et al., 2020).

The state of the art in educational evaluation research indicates a growing shift toward the use of more comprehensive inferential statistical methods capable of capturing the complexity inherent in educational data (Karimian et al., 2024). Two-Way Analysis of Variance (Two-Way ANOVA) has emerged as one of the most widely recommended approaches due to its ability to simultaneously examine the effects of two factors and their interaction on a response variable (Muzsnay et al., 2025). This approach provides a more robust analytical framework than one-way analyses or purely descriptive methods (Rutenberg et al., 2022).

Nevertheless, the application of Two-Way ANOVA in evaluating final semester examination outcomes in the Basic Statistics course, particularly within the Mathematics Education Study Program at Universitas Riau, remains relatively limited. At the international level, the growing movement toward learning analytics and evidence-based assessment has encouraged institutions to adopt multivariate and factorial approaches to better detect systematic performance differences across cohorts, class contexts, and student characteristics, especially when assessment data are used for curriculum review and quality assurance. However, many evaluation studies still rely on descriptive summaries or single-factor comparisons, which may overlook the simultaneous influence of instructional grouping (e.g., class sections) and student attributes (e.g., gender), as well as their interaction effects. This methodological gap is not merely institution-specific; it reflects a broader need for transparent, replicable, and scalable analytical frameworks that can be applied across higher education settings to support data-driven decision-making and improve comparability of learning outcome evaluations. Most existing studies continue to focus on single-factor effects or employ non-parametric approaches without an in-depth exploration of factor interactions. This condition highlights a research gap that warrants further investigation, especially in the context of data-driven academic evaluation in higher education.

Within the Mathematics Education Study Program at Universitas Riau, the evaluation of final semester examination scores in the Basic Statistics course is of particular importance, given that this course serves as a foundational prerequisite for advanced statistics courses and educational research. Understanding the dynamics of students' academic performance at the early stages of statistical learning can provide strategic insights for lecturers and program administrators in designing more adaptive and effective instructional practices (Al Hazaa et al., 2021).

The novelty of this study lies in the application of Two-Way ANOVA to comprehensively examine the dynamics of students' academic performance within the context of final semester evaluation (Premo et al., 2022)(Walker et al., 2020). Beyond a local application, this study aligns with international trends in data-driven educational evaluation by demonstrating a concise, interpretable factorial framework that can be replicated across institutions to examine how instructional grouping (class sections) and student characteristics (gender) jointly relate to achievement outcomes. The analysis provides evidence on whether observed score variability reflects systematic differences attributable to class-level implementation and student-level factors, rather than random fluctuation, thereby supporting quality assurance and equitable assessment practices. This research not only focuses on differences in learning outcomes but also explores the variation and interaction of academic factors influencing final examination scores. Such an approach is expected to yield a deeper, evidence-based understanding of academic performance patterns among mathematics education students.

Accordingly, this study is expected to contribute both theoretically and practically to the field of mathematics education evaluation. From a theoretical perspective, it enriches the methodological

discourse on student learning outcome analysis through the application of advanced inferential statistical approaches. From a practical standpoint, the findings may inform academic decision-making, instructional strategy improvement, and the enhancement of assessment quality in the Basic Statistics course within higher education.

METHODS

1. Research Design

This study employs an explanatory quantitative research design aimed at analyzing variations in students' academic performance within the context of final semester evaluation. A quantitative approach was selected because it enables objective measurement of academic achievement and facilitates inferential statistical hypothesis testing (Hwang et al., 2020) (Preston et al., 2020). The explanatory design is used to elucidate differences and patterns of variation in academic performance based on academic factors analyzed simultaneously (Heidari et al., 2023) (Kostromitina et al., 2025).

In addition, this study adopts a factorial design, which allows for the concurrent evaluation of multiple factors as well as the interaction effects among these factors on final examination scores (Sperling et al., 2024). A factorial design is considered particularly relevant in educational research, as students' academic performance is typically influenced by a combination of interacting factors rather than a single isolated variable. Accordingly, this design provides a more comprehensive analytical framework for explaining the dynamics of learning outcomes.

2. Participants and Research Context

The participants in this study were undergraduate students enrolled in the Mathematics Education Study Program who attended the Basic Statistics course during the academic semester under investigation. The Basic Statistics course is a compulsory subject designed to equip students with conceptual understanding and fundamental skills in descriptive and inferential statistics, which serve as a foundation for advanced coursework and research activities in mathematics education.

All students who had completed and sat for the final semester examination in this course were included as research subjects. The total sample size was N = 110 students. This approach was adopted to ensure comprehensive population coverage, thereby allowing the analytical results to reflect the actual academic conditions of students within the examined learning context. Therefore, the sampling technique used in this study was total sampling (census), where all eligible students in the population were included. The relatively homogeneous institutional and academic setting further enabled the analysis of variations in academic performance to be focused on relevant instructional factors, without substantial influence from differences in curricular structure.

3. Data and Research Variables

The primary data in this study consist of final semester examination scores of students from the Mathematics Education Study Program at Universitas Riau, which serve as the dependent variable and represent students' academic achievement. Final semester examination scores were selected because they constitute a form of summative assessment designed to comprehensively measure students' competency attainment after the completion of all instructional content in the Basic Statistics course, thereby providing an objective representation of final learning outcomes.

The final semester examination (FSE) was constructed based on the course learning outcomes and the official course syllabus. The examination covered key competencies in descriptive statistics, probability concepts, and introductory inferential statistics. Importantly, the FSE was administered as a standardized assessment across Class A, Class B, and Class C, meaning that the test blueprint, item set, time allocation, and scoring rubric were consistent across classes to ensure comparability of scores.

This study involves two categorical independent variables, namely class and gender. The class variable comprises three levels: Class A, Class B, and Class C, while the gender variable consists of two levels: male and female. These variables were selected based on their relevance to the instructional context and their potential influence on students' academic achievement. The factorial structure of the data allows for the examination of the main effects of each independent variable, as well as the interaction effect between class and gender on final semester examination scores (Hall et al., 2023).

Data were collected after the completion of the FSE through documentation of official score records provided by the course administration. Prior to analysis, the dataset was screened for completeness and accuracy (e.g., missing values and duplicate entries). Additionally, potential outliers were examined using boxplots and standardized residual diagnostics. Although one observation appeared as a potential outlier in the class-level boxplot visualization, no data points were removed from the analysis, because the score remained within the plausible range of assessment outcomes and did not indicate data entry error. All observations were retained to preserve the integrity of the population-based (total sampling) dataset.

Through this factorial approach, the study not only identifies differences in students' academic performance across classes and gender groups but also elucidates how the combination of these two variables simultaneously contributes to variations in learning outcomes. This analytical approach provides a deeper, evidence-based understanding of the dynamics of students' academic achievement within the context of final semester evaluation and supports more comprehensive result interpretation in higher education research (Renaldo et al., 2023).

4. Two-Way Analysis of Variance (Two-Way ANOVA)

Two-Way Analysis of Variance (Two-Way ANOVA) is a parametric statistical method used to test differences in the means of a quantitative response variable based on two categorical factors simultaneously, while also evaluating the interaction effect between these factors (Lim et al., 2022). This method provides a more comprehensive analytical framework than one-way ANOVA, as it is capable of capturing multidimensional variation dynamics within the data (Hsu et al., 2023).

In educational research, students' academic achievement is not shaped by a single factor, but rather by a combination of academic variables and learning-context factors that interact with one another (Stefanescu & Trandafir, 2024). Two-Way ANOVA enables researchers to disentangle the main effects of each factor and to identify interaction effects, which occur when the effect of one factor depends on the level of another factor (Vijayaragunathan & Srinivasan, 2020). Consequently, this approach is highly relevant for analyzing learning evaluation outcomes, including final semester examination scores.

Mathematically, the Two-Way ANOVA model with interaction can be formulated as follows (Rayarao, 2025):

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk} \quad (1)$$

where:

- Y_{ijk}: academic performance score (final semester examination score) of the k-th student under the combination of factor A at level $i$and factor B at level j
- μ: overall mean (grand mean)
- α_i: main effect of factor A
- β_j: main effect of factor B
- $(\alpha\beta)_{ij}$: interaction effect between factors A and B
- ε_{ijk}: random error term assumed to be normally distributed with a mean of zero and constant variance

This model allows for the simultaneous testing of hypotheses related to two factors and their interaction, thereby improving analytical efficiency and enhancing the depth of result interpretation (Durell et al., 2025).

Hypothesis Testing

Two-Way ANOVA is employed to test three primary hypotheses:

1. There is no difference in mean academic performance across the levels of factor A.
2. There is no difference in mean academic performance across the levels of factor B.
3. There is no interaction effect between factor A and factor B on academic performance.

Rejection of any null hypothesis indicates that the tested factor or interaction contributes significantly to the variation in academic performance.

Variance Decomposition

The total variation in the data under the Two-Way ANOVA framework can be decomposed as follows:

$$SS_{\text{Total}} = SS_A + SS_B + SS_{AB} + SS_{\text{Error}} \quad (2)$$

where:

- SS_A: sum of squares due to factor A
- SS_B: sum of squares due to factor B
- SS_{AB}: sum of squares due to the interaction between factors A and B
- SS_{Error}: error sum of squares (variation unexplained by the model)

This decomposition forms the basis for calculating mean squares and the F-test statistics.

F-Test Statistics

Mean squares are calculated as:

$$MS = \frac{SS}{df} \tag{3}$$

The F-test statistics for each source of variation are given by:

$$F_A = \frac{MS_A}{MS_{\text{Error}}}, F_B = \frac{MS_B}{MS_{\text{Error}}}, F_{AB} = \frac{MS_{AB}}{MS_{\text{Error}}} \tag{4}$$

The computed F-values are subsequently compared with the critical values from the F-distribution or evaluated using significance levels (p-values).

Table 1. Two-Way ANOVA Table

Source	SS	df	MS	F	p-value
Factor A	SS_A	a-1	MS_A	F_A	p_A
Factor B	SS_B:	b-1	MS_B	F_B	p_A
A×B Interaction	SS_{AB}	(a-1)(b-1)	MS_{AB}	F_{AB}	p_A
Error	SS_{Error}	N-ab	MS_{Error}		
Total	SST	N-1			

RESULTS AND DISCUSSION

a. Distribution and Differences in Assessment Scores by Class and Gender

To explore the distributional patterns of assessment scores across class and gender groups, data visualization was conducted using boxplots. This approach was employed to illustrate differences in medians, data dispersion, and potential variability among groups without relying initially on inferential test results. Such visualizations provide preliminary insights into the possible presence of main effects and interaction effects between class and gender factors on assessment scores.

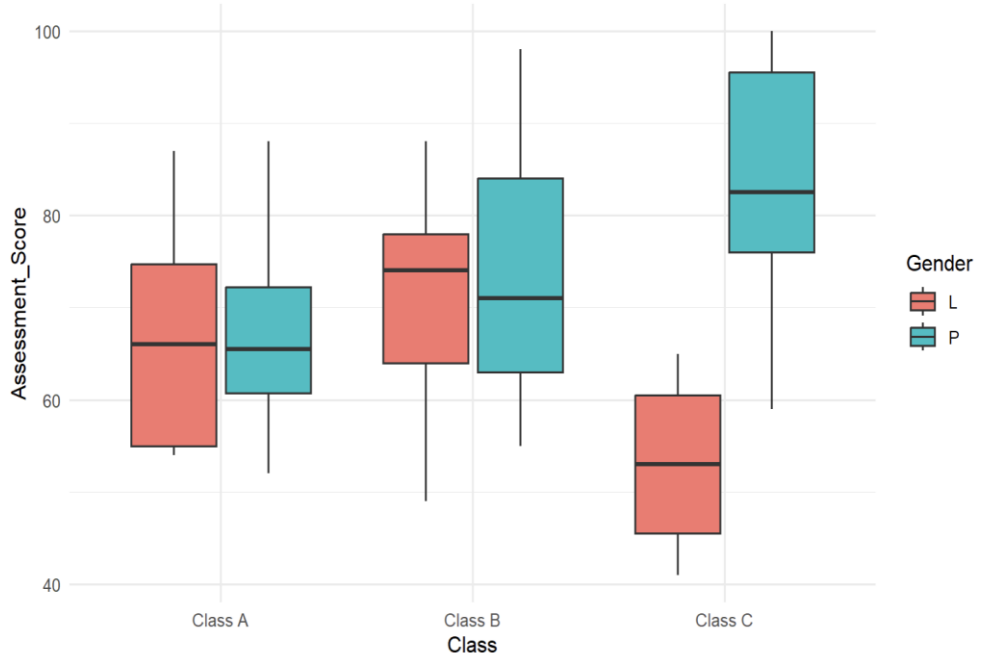

Figure 1. Distribution of Assessment Scores by Class and Gender

Based on Figure 1, the distribution of assessment scores varies across classes and gender groups. Overall, the median scores in Class C tend to be higher than those in Class A and Class B, particularly among the female group. This pattern indicates the presence of differences in academic achievement across classes.

From a gender perspective, female students consistently exhibit higher median scores than male students across all classes. This difference is most pronounced in Class C, where the gap between the medians of the two gender groups is relatively larger compared to the other classes. In Class A and Class B, gender-based differences remain observable but with more moderate magnitude.

From a pedagogical perspective, the tendency for female students to obtain higher scores may be associated with differences in learning strategies and self-regulation that are frequently reported in the educational literature, such as greater consistency in completing practice tasks, sustained persistence throughout the learning process, and more structured time management. In a Basic Statistics course, which demands procedural accuracy, repeated practice, and the ability to read and interpret data, stable study habits and disciplined practice can provide an advantage in examination performance. In addition, several educational studies suggest that female students often demonstrate more consistent engagement in classroom activities and formative assignments, which may ultimately contribute to stronger summative outcomes. However, because this study did not directly measure variables such as motivation, learning strategies, or engagement, this explanation should be interpreted as a plausible pedagogical interpretation and as a basis for future research, for example by incorporating indicators such as attendance, coursework/quiz scores, or self-regulated learning measures to test the underlying mechanisms more specifically.

In addition, score dispersion varies across combinations of class and gender. Female students in Class B and Class C demonstrate wider score ranges, suggesting greater heterogeneity in learning achievement within these groups. The misalignment of median scores between male and female students across classes provides preliminary evidence of a potential interaction between class and gender factors on assessment scores. This indication, however, requires confirmation through inferential analysis, such as Two-Way Analysis of Variance (Two-Way ANOVA).

b. Interaction Effects between Class and Gender on Assessment Scores

To identify the presence of an interaction effect between class and gender on assessment scores, an interaction plot was constructed to display the mean assessment scores for each factor combination. This visualization serves as an initial diagnostic tool to evaluate whether the effect of gender on assessment scores is consistent across classes or varies depending on specific class levels.

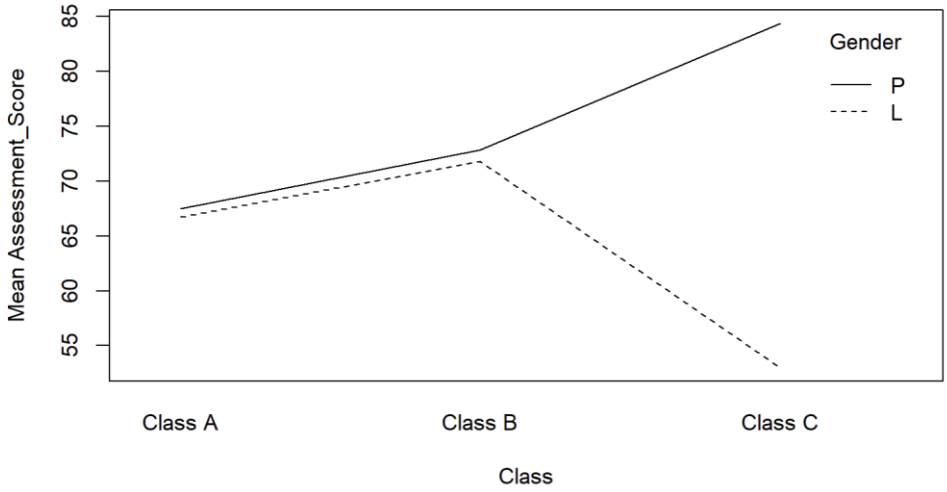

Figure 2. Interaction Plot of Mean Assessment Scores by Class and Gender

Based on Figure 2, a clear interaction pattern between class and gender on mean assessment scores is observed. Among female students, assessment scores exhibit a consistent increasing trend from Class A to Class C. In contrast, among male students, assessment scores increase from Class A to Class B but show a pronounced decline in Class C.

The non-parallel and intersecting lines between the two gender groups, particularly during the transition from Class B to Class C, indicate the presence of a substantial interaction effect between class and gender. These findings suggest that the influence of gender on assessment scores is not uniform across all classes but instead depends on the specific class level attended. Accordingly, this visualization provides preliminary supporting evidence that the combined effects of class and gender simultaneously contribute to variations in assessment scores. This observation, however, requires confirmation through formal inferential testing using Two-Way Analysis of Variance (Two-Way ANOVA).

c. Classical Assumptions

Prior to conducting the analysis of variance, statistical assumption testing was performed to ensure that the data met the prerequisites for the application of Analysis of Variance (ANOVA). The assumptions examined included the normality of the residual distribution and the homogeneity of variances across groups. Satisfying these assumptions is essential to ensure the validity of the statistical inferences derived from the ANOVA results.

Table 2. Results of Assumption Tests for ANOVA

Assumptions	Test	P-value	Decision
Normal	Shapiro-Wilk	0.1854	Fulfilled
Homogenitas	Levane test	0.6360	Fulfilled

Based on the results presented in Table 2, the Shapiro–Wilk test yielded a p-value of 0.1854, which exceeds the significance level of 0.05. This result indicates that the data do not exhibit a significant deviation from normality, thereby satisfying the normality assumption. Furthermore, Levene's test produced a p-value of 0.6360, which is also greater than 0.05. This finding confirms that the variances across groups are homogeneous, indicating that the assumption of homogeneity of variances is met.

With both the normality and homogeneity of variance assumptions satisfied, the data meet all prerequisites for conducting Analysis of Variance (ANOVA). Consequently, subsequent ANOVA procedures can be validly performed to examine the effects of the investigated factors and their interaction on the response variable.

d. Two-Way Analysis of Variance (Two-Way ANOVA)

To examine the effects of class, gender, and the interaction between these two factors on assessment scores, a Two-Way Analysis of Variance (Two-Way ANOVA) was conducted. This analysis was performed after all statistical assumptions namely normality and homogeneity of variances were

confirmed to be satisfied. Two-Way ANOVA enables the simultaneous testing of the main effects of each factor as well as their interaction effects on the dependent variable.

Table 3. Two-Way Analysis of Variance (ANOVA) for Assessment Scores

Source of Variation	df	Sum Of Squares	Mean Square	F-value	p-value
Class	2	3151	1575.3	12.191	1.75e-05 ***
Gender	1	901	901.1	6.974	0.00954 **
Class x Gender	2	2583	1291.3	9.993	0.000107 ***
Residual	104	13439	129.2		

Based on the Two-Way ANOVA results reported in Table 3, the class factor has a statistically significant effect on assessment scores ($F = 12.191; p = 1.75e - 05$). This indicates that the mean scores differ meaningfully across classes (Class A, B, and C), suggesting that "class" as a learning context contributes to variation in students' performance. Gender also shows a statistically significant effect on assessment scores ($F = 6.974; p = 0.00954$). In other words, when considered overall, there is a significant difference in mean scores between male and female students.

Most importantly, the class and gender interaction is statistically significant ($F = 9.993; p = 0.000107$). This implies that the effect of gender on scores is not uniform across classes; put differently, the performance gap between male and female students depends on the specific class. Because the interaction is significant, interpretation should not rely solely on the main effects, but should emphasize that the gender-related pattern varies by class. Practically, this is consistent with interaction plots that typically show non-parallel or crossing lines, indicating that in certain classes the gender gap may widen, narrow, or even reverse.

Overall, these findings confirm that variation in assessment scores is influenced by class, gender, and, crucially, by their combined effect (the interaction). Therefore, an appropriate next step is to examine differences at the level of factor combinations (e.g., comparing genders within each class, or comparing classes within each gender) using follow-up analyses such as simple effects tests or relevant post hoc procedures.

Given the significant class and gender interaction, the results also point to class-specific learning dynamics that shape achievement for male and female students. In other words, gender differences in scores do not appear to be stable and uniform; instead, they may strengthen or weaken depending on the class context, such as differences in instructional implementation, the intensity of practice and feedback, the learning climate, or the composition and study strategies of students within a class. Pedagogically, this matters because it suggests that improvement efforts should not rely on a "one-size-fits-all" approach across classes, but should be adapted to the characteristics of each class and student group. Accordingly, follow-up analyses such as class- or gender-specific simple effects, together with an examination of process-related factors, are relevant for explaining why gender differences vary across classes and for targeting instructional improvements more precisely.

e. Post Hoc Mean Comparisons Using Tukey's Honestly Significant Difference (Tukey HSD)

Following the Two-Way Analysis of Variance (Two-Way ANOVA), which revealed significant effects of class and gender on assessment scores, a post hoc analysis was conducted using Tukey's Honestly Significant Difference (Tukey HSD) method. The Tukey HSD test was selected because it is appropriate for performing pairwise comparisons among groups while simultaneously controlling the family-wise Type I error rate, thereby allowing for more reliable interpretation of mean differences across groups.

Table 4. Results of the Tukey HSD Test for Pairwise Comparisons among Classes

Class	diff	lwr	upr	p-value
Class B-Class A	5.266667	- 0.9428671	11.47620	0.113250
Class C-Class A	13.343137	6.8792800	19.80699	0.0000102
Class C-Class B	8.076471	1.7715467	14.38139	0.0082034

Based on the Tukey HSD results presented in Table 4, the comparison between Class C and Class A shows a statistically significant difference in mean assessment scores (p-value = 0.0000102). This indicates that students in Class C achieved significantly higher assessment scores than those in

Class A. In addition, the comparison between Class C and Class B is also statistically significant (p-value = 0.008203), suggesting that Class C likewise outperformed Class B.

In contrast, the comparison between Class B and Class A does not reveal a statistically significant difference (p-value = 0.11325). Collectively, these findings suggest that class-related differences in assessment scores are primarily driven by the superior performance of Class C, whereas Classes A and B do not differ significantly after adjustment for multiple comparisons.

Table 5. Tukey HSD Test Results for Comparison of Assessment Scores Based on Gender

Gender	diff	lwr	upr	P-value
P-L	7.225779	1.756968	12.69459	0,0101051

The Tukey HSD results presented in Table 5 indicate a statistically significant difference in mean assessment scores between female and male students (p = 0.0101). The positive mean difference (P–L = 7.226) suggests that female students tend to achieve higher assessment scores than male students overall.

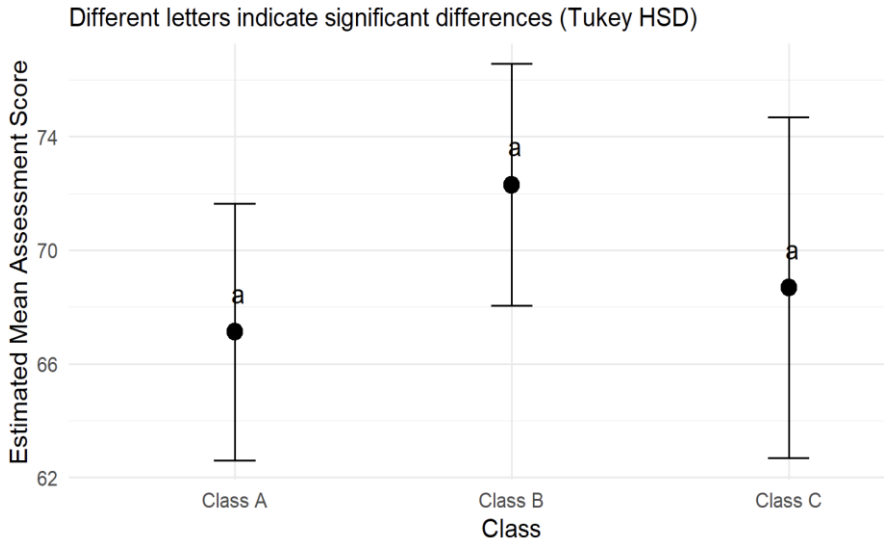

Figure 3. Tukey HSD Multiple Comparisons of Mean Assessment Scores Across Classes

Figure 3 presents the Tukey HSD multiple-comparison visualization of mean assessment scores across Class A, Class B, and Class C. The black dots represent the estimated mean scores for each class, while the vertical error bars reflect the uncertainty around these estimates (e.g., confidence intervals). In this figure, the compact letter display ("a", "b", etc.) summarizes the Tukey HSD grouping rule: classes sharing the same letter are not statistically different after adjustment for multiple comparisons, whereas classes with different letters indicate statistically significant differences.

Based on Figure 3, all classes (Class A, Class B, and Class C) share the same grouping letter ("a"). This indicates that, according to the Tukey HSD grouping represented in the figure, no pairwise differences among the three class means are statistically significant once the Tukey adjustment is applied. Although the plotted means suggest that Class B and Class C tend to have higher average scores than Class A, the Tukey procedure evaluates these differences relative to the within-group variability and the sample sizes, and then applies a multiple-comparison correction. Under these conditions, the observed mean gaps may not be sufficiently large to exceed the Tukey critical threshold, resulting in a single homogeneous group ("a") across all classes.

The fact that all classes receive the same letter can be explained by the relationship between the figure and the numerical Tukey outputs. Specifically, a uniform "a" labeling occurs when the Tukey-adjusted p-values for all pairwise class comparisons exceed the chosen significance level (typically 0.05), meaning that none of the class-to-class differences are statistically reliable after controlling the family-wise error rate. In practice, this can occur when (i) the dispersion of scores within classes is relatively large (wide error bars), (ii) the sample size per class is not sufficiently large to detect moderate mean differences, or (iii) the multiple-comparison correction reduces sensitivity compared with unadjusted tests. Therefore, the identical letters do not mean that the class means are exactly equal, but

rather that the evidence is insufficient to declare statistically significant differences among classes under the Tukey HSD criterion.

However, it is essential that the compact letter display in Figure 3 be consistent with the corresponding Tukey HSD table. If the Tukey table reports statistically significant differences for certain pairs (e.g., Class C–Class A or Class C–Class B), then Figure 3 should reflect that by assigning different letters to the significantly different groups. Conversely, if the table shows non-significant adjusted p-values for all pairs, then the "all a" grouping in Figure 3 is appropriate. Accordingly, to ensure transparent and error-free reporting, Figure 3 should be generated directly from the same Tukey HSD output used to produce the table, so that the letter groupings and pairwise p-values are fully aligned.

f. Implications

The findings of this study indicate that students' academic achievement in the context of final semester evaluation is significantly influenced by class and gender as main effects, while the interaction between these two factors does not exhibit a statistically significant effect. This pattern suggests that variations in students' academic outcomes are shaped through the independent contributions of each factor, without a strong structural dependency between class and gender. From a theoretical perspective, these results reinforce the view that, within relatively homogeneous learning contexts, additive mechanisms of influence are more dominant than interactive mechanisms in explaining differences in learning outcomes (Wrigley-Asante et al., 2023).

Conceptually, the non-significant interaction effect between class and gender indicates that gender-based differences in academic achievement tend to be consistent across classes, and conversely, class-based differences apply relatively uniformly to both gender groups (Gil et al., 2021). This finding provides empirical support for the use of factorial analytical models that emphasize the interpretation of main effects in higher education evaluation research. Accordingly, this study contributes to strengthening the methodological framework for investigating academic achievement, particularly in courses characterized by standardized curricular structures and uniform assessment systems.

From a practical standpoint, the significant effect of class on final examination scores suggests the presence of variability in classroom-level learning dynamics, despite the use of identical curricula and instructional materials (Kostromitina et al., 2025b). This underscores the importance of continuous evaluation and reflection on instructional practices within each class, including pedagogical strategies, classroom management, and the intensity of lecturer–student interactions. By ensuring greater consistency in instructional implementation, higher education institutions may promote more equitable learning quality and academic achievement among students.

Furthermore, the significant influence of gender on academic achievement carries important implications for the design of learning environments that are more inclusive and responsive to student diversity. Although gender-based class differentiation is not warranted, these findings highlight the need for instructional and assessment strategies that accommodate differences in learning styles, engagement patterns, and participation. Overall, this study provides a strong empirical foundation for data-driven academic decision-making and opens avenues for future research to explore additional factors that may comprehensively influence students' academic achievement.

CONCLUSION

This study examined differences in students' final examination performance in Basic Statistics across class sections and gender. The findings highlight that learning outcomes are not evenly distributed across classes, with Class C consistently achieving higher results than Class A. This pattern suggests that classroom-level factors such as instructional implementation, learning support, and assessment preparation practices may shape students' achievement beyond individual ability alone.

From a practical perspective, the results provide a clear message for educators and program administrators: improvement efforts should prioritize reducing achievement gaps between classes. Specifically, strategies that appear to work effectively in Class C should be identified and adapted for Class A. Recommended actions include standardizing key teaching elements across sections, such as pacing, practice intensity, and feedback routines; implementing targeted academic support for students in lower-performing classes through structured tutorials, additional practice sessions, or formative

quizzes; and monitoring learning progress throughout the semester to allow early intervention rather than relying solely on end-of-semester outcomes.

In addition, the persistent difference observed between male and female students indicates the need for inclusive instructional support that strengthens engagement and learning strategies for all students. Program-level monitoring of performance patterns, combined with reflective teaching evaluation, can help ensure that assessment outcomes are used not only for grading but also for continuous instructional improvement. Overall, this study reinforces the value of evidence-informed academic decision-making and supports more equitable and effective learning evaluation in higher education, particularly in foundational courses such as Basic Statistics.

ACKNOWLEDGEMENT

The authors would like to express their sincere gratitude to the Mathematics Education Study Program, University of Riau, for providing academic support and facilitating the implementation of this research. The support and cooperation received throughout the research process are gratefully acknowledged.

REFERENCES

Al Hazaa, K., Abdel-Salam, A.-S. G., Ismail, R., Johnson, C., Al-Tameemi, R. A. N., Romanowski, M. H., BenSaid, A., Rhouma, M. B. H., & Elatawneh, A. (2021a). The effects of attendance and high school GPA on student performance in first-year undergraduate courses. *Cogent Education*, *8*(1), 1956857.

Al Hazaa, K., Abdel-Salam, A.-S. G., Ismail, R., Johnson, C., Al-Tameemi, R. A. N., Romanowski, M. H., BenSaid, A., Rhouma, M. B. H., & Elatawneh, A. (2021b). The effects of attendance and high school GPA on student performance in first-year undergraduate courses. *Cogent Education*, *8*(1), 1956857.

Ali, Q. I. (2024). Towards more effective summative assessment in OBE: a new framework integrating direct measurements and technology. *Discover Education*, *3*(1), 107.

Durell, L., Mastin, N., Lovin, L. M., Steele IV, W. B., Brooks, B. W., & Hering, A. S. (2025). Balanced Two-Way Functional ANOVA: A Case Study for Toxicological Photolocomotor Response Studies. *Journal of Agricultural, Biological and Environmental Statistics*, 1–30.

Gil, P. D., da Cruz Martins, S., Moro, S., & Costa, J. M. (2021). A data-driven approach to predict first-year students' academic success in higher education institutions. *Education and Information Technologies*, *26*(2), 2165–2190.

Guo, P., Saab, N., Post, L. S., & Admiraal, W. (2020). A review of project-based learning in higher education: Student outcomes and measures. *International Journal of Educational Research*, *102*, 101586.

Hall, C., Dahl-Leonard, K., Cho, E., Solari, E. J., Capin, P., Conner, C. L., Henry, A. R., Cook, L., Hayes, L., & Vargas, I. (2023). Forty years of reading intervention research for elementary students with or at risk for dyslexia: A systematic review and meta-analysis. *Reading Research Quarterly*, *58*(2), 285–312.

Hamilton, D., McKechnie, J., Edgerton, E., & Wilson, C. (2021). Immersive virtual reality as a pedagogical tool in education: a systematic literature review of quantitative learning outcomes and experimental design. *Journal of Computers in Education*, *8*(1), 1–32.

Heidari, H., Beni, Z. H. mirzaee, & Deris, F. (2023). Using Kern model to design, implement, and evaluate an infection control program for improving knowledge and performance among undergraduate nursing students: a mixed methods study. *BMC Medical Education*, *23*(1), 795.

Hsu, T.-C., Chen, W.-L., & Hwang, G.-J. (2023). Impacts of interactions between peer assessment and learning styles on students' mobile learning achievements and motivations in vocational design certification courses. *Interactive Learning Environments*, *31*(3), 1351–1363.

Hwang, G.-J., Sung, H.-Y., Chang, S.-C., & Huang, X.-C. (2020). A fuzzy expert system-based adaptive learning approach to improving students' learning performances by considering affective and cognitive factors. *Computers and Education: Artificial Intelligence*, *1*, 100003.

Ismail, S. M., Rahul, D. R., Patra, I., & Rezvani, E. (2022). Formative vs. summative assessment: impacts on academic motivation, attitude toward learning, test anxiety, and self-regulation skill. *Language Testing in Asia*, *12*(1), 40.

Karimian, Z., Mokarram, P., & Zarifsanaiey, N. (2024). Comparison of the teaching clinical biochemistry in face-to-face and the flex-flipped classroom to medical and dental students: a quasi-experimental study from IRAN. *BMC Medical Education*, *24*(1), 137.

Kostromitina, M., Naismith, B., Burstein, J., & Plonsky, L. (2025a). Beyond GPA and language proficiency: A systematic literature review of international students' academic success factors. *Review of Education*, *13*(2), e70089.

Kostromitina, M., Naismith, B., Burstein, J., & Plonsky, L. (2025b). Beyond GPA and language proficiency: A systematic literature review of international students' academic success factors. *Review of Education*, *13*(2), e70089.

Lim, J., Ko, H., Park, J., & Ihm, J. (2022). Effect of active learning and online discussions on the academic performances of dental students. *BMC Medical Education*, *22*(1), 312.

Muzsnay, A., Szabó, C., Zámbó, C., Szabó, G., & Szeibert, J. (2025). Retrieval Practice—A Tool to Narrow the Achievement Gap in Learning Higher Mathematics. *International Journal of Science and Mathematics Education*, 1–27.

Premo, J., Wyatt, B. N., Horn, M., & Wilson-Ashworth, H. (2022). Which group dynamics matter: social predictors of student achievement in team-based undergraduate science classrooms. *CBE—Life Sciences Education*, *21*(3), ar51.

Preston, R., Gratani, M., Owens, K., Roche, P., Zimanyi, M., & Malau-Aduli, B. (2020). Exploring the impact of assessment on medical students' learning. *Assessment & Evaluation in Higher Education*, *45*(1), 109–124.

Rabattu, P., Debarnot, U., & Hoyek, N. (2023). Exploring the impact of interactive movement-based anatomy learning in real classroom setting among kinesiology students. *Anatomical Sciences Education*, *16*(1), 148–156.

Rayarao, S. R. (2025). Two-way Analysis of Variance: A Comprehensive Review of Theory, Applications, and Statistical Methodology. *Authorea Preprints*.

Renaldo, N., Sevendy, T., & Purnama, I. (2023). Improving accounting students' statistical understanding of 2-way anova through a case study of indonesian coffee exports. *Reflection: Education and Pedagogical Insights*, *1*(1), 13–19.

Rutenberg, I., Ainscough, L., Colthorpe, K., & Langfield, T. (2022a). The anatomy of agency: Improving academic performance in first-year university students. *Anatomical Sciences Education*, *15*(6), 1018–1031.

Rutenberg, I., Ainscough, L., Colthorpe, K., & Langfield, T. (2022b). The anatomy of agency: Improving academic performance in first-year university students. *Anatomical Sciences Education*, *15*(6), 1018–1031.

Sperling, J., Mburi, M., Gray, M., Schmid, L., & Saterbak, A. (2024). Effects of a first-year undergraduate engineering design course: Survey study of implications for student self-efficacy and professional skills, with focus on gender/sex and race/ethnicity. *International Journal of STEM Education*, *11*(1), 8.

Stefanescu, I. O., & Trandafir, R. (2024). Participation in Lifelong Learning Depending on Economic Development and Educational Attainment level: A Statistical Approach Using Two-Way ANOVA. *Economics and Applied Informatics*, *3*, 256–264.

Uzun, Y., Suraworachet, W., Zhou, Q., Gauthier, A., & Cukurova, M. (2025). Engagement with analytics feedback and its relationship to self-regulated learning competence and course performance. *International Journal of Educational Technology in Higher Education*, *22*(1), 17.

Vijayaragunathan, R., & Srinivasan, M. R. (2020). Bayes factors for comparison of two-way ANOVA models. *Journal of Statistical Theory and Applications*, *19*(4), 540–546.

Walker, E. R., Lang, D. L., Caruso, B. A., & Salas-Hernández, L. (2020). Role of team dynamics in the learning process: a mixed-methods evaluation of a modified team-based learning approach in

a behavioral research methods course. *Advances in Health Sciences Education*, *25*(2), 383–399.

Wrigley-Asante, C., Ackah, C. G., & Frimpong, L. K. (2023). Gender differences in academic performance of students studying Science Technology Engineering and Mathematics (STEM) subjects at the University of Ghana. *SN Social Sciences*, *3*(1), 12.

Yu, Z., Yu, L., Xu, Q., Xu, W., & Wu, P. (2022). Effects of mobile learning technologies and social media tools on student engagement and learning outcomes of English learning. *Technology, Pedagogy and Education*, *31*(3), 381–398.