European Journal of Education and Pedagogy
Vol 6 | Issue 4 | July 2025
ISSN 2736-4534

RESEARCH ARTICLE

A Ten-Year Study of Sex-Based Differences in Cognitive Development in University Essay Writing

Oscar E. Quirós [1,*] and Georgina Morera-Quesada [2]

ABSTRACT

This ten-year study analyzes sex differences and cognitive development in essay writing among 635 university students at the University of Costa Rica's Golfito campus, based on 1,863 essays from Theatre and Film Appreciation courses. Using a fixed rubric and a single evaluator, the study tracked essay performance across pre-, during-, and post-COVID periods (2015–2024). Female students consistently outperformed male students in average essay grades, with the gap widening during the pandemic. Both sexes, however, showed similar improvement trajectories throughout the semester. Students with lower initial scores improved more over time, suggesting motivational and instructional effects not captured by sex alone. A marginal interaction between sex, initial performance, and the COVID period appeared only in isolated models and did not persist longitudinally. These findings highlight the effectiveness of essay-based assessment in promoting higher-order thinking and tracking learning progression. The study also emphasizes the importance of sustained, context-aware evaluation to reveal both persistent inequalities and resilient cognitive growth.

Keywords: Cognitive development, essay writing, formative assessment, longitudinal study.

Submitted: May 27, 2025

Published: July 15, 2025

10.24018/ejedu.2025.6.4.971

[1] *Professor, Theatre and Film; General Studies, University of Costa Rica, Golfito.*
[2] *Professor, General Studies, University of Costa Rica, Golfito.*

* ***Corresponding Author:***
e-mail: oscar.quirosruiz@ucr.ac.cr

1. Introduction

Academic assessment plays a central role in identifying how and what students are learning across different academic contexts. By systematically evaluating student performance, educators can refine teaching methods and instructional strategies. In university-level arts education, where expressive, creative, and analytical abilities converge, evaluation poses unique challenges due to its multidimensional nature (Henry *et al*., 2005). Analytical essays, in particular, are valuable tools for assessing critical thinking, argumentation, and cognitive depth. As stated in Bloom's Taxonomy and later revisions (Anderson, 2005; Bloom *et al*., 1956), cognitive abilities such as comprehension, analysis, evaluation, and synthesis, are highly effective ways for evaluating and stimulating the development of students' higher level capabilities. Writing essays, therefore challenge and serve to stimulate students cognitive competence at different levels.

Short analytical essays are one of the optimal means of assessing cognitive performance in students because they require a wide range set of higher-level skills. In particular, they require students to: (1) understand key concepts and evidence (Comprehension); (2) apply those concepts to a specific hypothesis or argument (Application); (3) analyze the logical relationships between ideas and evidence (Analysis); (4) evaluate the strength and relevance of sources or reasoning (Evaluation); and (5) create a coherent argument or new insight through structured writing (Creation). In doing so, students must assess the strength of each piece of evidence, identify implicit assumptions, construct logical and causal relationships, and synthesize a cohesive conclusion. In addition, strong written communication and metacognitive awareness are also essential, ensuring the essay is clearly structured so that evidence, by means of explanations, supports a stated hypothesis.

This study analyzes essay performance across two arts courses, *Theatre Workshop* and *Film Appreciation*, taught at the Golfito campus of the University of Costa Rica (UCR) between 2015 and 2024. Our ten-year dataset includes 635 students and 1,863 essays, all graded using a fixed rubric by a single evaluator. The central aim is

to explore performance differences by sex across three time periods: before, during, and after the COVID-19 pandemic. This focus allows for a detailed investigation of how sex experiences, including possible differential access to technology, emotional burdens, and role expectations, may have affected academic persistence and essay quality.

The pandemic posed unprecedented challenges to higher education worldwide. The transition to online learning magnified existing inequalities, especially in rural and socioeconomically vulnerable regions such as Southwestern Costa Rica (Vargas *et al*., 2023), where the Golfito campus is located. National data indicate that academic performance declined in many areas during the pandemic, particularly among students with limited access to digital tools (Barquero *et al*., 2021). Several international studies confirm that sex played a significant role in mediating these effects: women often faced increased domestic responsibilities, emotional distress, and reduced capacity for academic engagement (Bertoletti *et al*., 2023; Bratti & Lippo, 2022; Zarowski *et al*., 2024). Alternatively, male students were more likely to exhibit performance variability linked to adaptation difficulties or disengagement with online platforms (Alshaibani *et al*., 2023; Kaba *et al*., 2024), a condition that we perceived on our campus.

Furthermore, the effects of the pandemic on academic performance were not always linear or uniform over time. For instance, some institutions saw improved grades during 2020 due to lenient attitudes towards grading, lower content coverage, or asynchronous learning models (Whitelaw *et al*., 2024), while others documented performance declines that disproportionately affected women or low-income students (Ferrer *et al*., 2023; Montolio & Taberner, 2018). In the context of essay-based evaluation, sex differences have also been noted in argumentation quality and peer feedback, with female students generally exhibiting stronger justification and cohesion in their writing (Noroozi *et al*., 2023). The latter is a prevalent issue in our campus.

This study contributes to a growing body of research on pre, during, and post-pandemic academic recovery and sex-based disparities in higher education. By focusing on a relatively underrepresented region and using essays as an evaluative instrument, it seeks to shed light on how sex and period-specific context influence not only persistence after academic setbacks, but also the depth and quality of student writing. This study draws on cognitive development frameworks in higher education, particularly Bloom's revised taxonomy (Anderson *et al*., 2001) and the notion of cumulative knowledge construction. These models inform our interpretation of students' essay performance and cognitive trajectories over time. By measuring higher cognitive competences for a longer period of time, we can have a more substantive assessment and evaluation of performance by both sex groups.

2. Method

2.1. Participants and Data Collection

Our dataset consists of grades from two art classes (Theatre and Cinema) in which students were required to write up to four, short 5-paragraph, analytical essays about different plays and cinema pieces. In both classes, we are not just asking students to report evidence, we are asking them to connect, justify, and evaluate. That places this task firmly in the realm of analytical and critical reasoning, rather than rote learning or descriptive writing. The dataset includes all essay grades from March 2015 and December 2024, covering a total of 635 students and 1.863 essays.

All essays were graded by only one of the authors and using the same criteria throughout the entire ten-year span (Table I). While all essays were graded by a single evaluator, this approach ensured high internal consistency across the ten-year period. The evaluator applied a detailed and stable rubric throughout, minimizing inter-cohort variation and maintaining uniformity in scoring standards. Although this design limits inter-rater reliability, it was deemed appropriate for this context given the longitudinal scope and consistent rubric application. Nevertheless, we acknowledge this as a methodological limitation. Previous studies have shown that a consistent rater using a structured rubric can yield stable and valid results in longitudinal educational assessments (Jonsson & Svingby, 2007). Every essay includes a formative assessment feedback about the deficiencies and flaws, so that students can learn and improve upon it. The intention is that the following essay would be better than the previous one. Future research could enhance this design by incorporating inter-rater calibration or independent double-coding to confirm scoring reliability.

2.2. Variables and Measures

We divided the dataset in two, one with just grade averages, and the second one with each individual grade from all 4 essays. We analyzed the dataset containing student essay grades from both classes, which included student ID, semester, year, course, sex, and average essay grade. We use the term "gender" throughout this study to refer to binary categories (male/female) as recorded in institutional enrollment data, which may reflect sex assigned at birth or gender identity depending on student disclosure. We retained only the first recorded course attempt per student, ensuring that each individual contributed a single observation. Although the sample included an unequal number of male and female students, this imbalance does not bias the analysis. All models explicitly included sex as a fixed effect and examined sex-specific effects and interactions. In particular, the linear and mixed-effects models adjusted for any differences in group size, ensuring that comparisons reflect true performance differences rather than sampling imbalance. The Poisson regression model for essay count similarly estimated mean values separately for each sex, and no significant differences were found in submission behavior. Additionally, confidence intervals and p-values were computed using robust methods, including parametric bootstrapping for the mixed model. Together, these strategies mitigate concerns about statistical bias due to sex imbalance.

2.3. Analytic Strategy

We categorized the academic years into three periods: pre-covid (2015–2019), covid (2020–2021), and post-covid

TABLE I: Essay Grading Criteria

Evaluation criterion	Description of the criterion	Max points	Points awarded
1. Clearly stated and well-formulated thesis	The main hypothesis is clearly stated, specific, and relevant to the assigned topic.	3,00	2,50
2. Relevant evidence and argumentative analysis	The evidence is clearly explained, relevant to the thesis, and justified through reasoning. The link between evidence and argument is analytical, not merely descriptive.	3,00	2,00
3. Logical structure and organization	The essay presents a clear structure (introduction, body, conclusion), with well-sequenced paragraphs and coherent progression.	2,00	2,00
4. Clarity and correctness in writing	The text is clearly written, grammatically correct, and free of spelling or punctuation errors. The style is semantically fluent and precise.	2,00	1,75
		10 pts	**8,25**

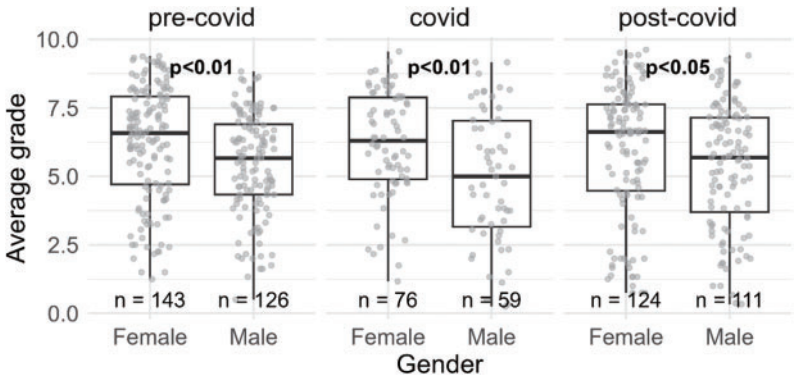

Fig. 1. Distribution of average essay grades by gender/sex across three periods, where each dot represents a student's first-attempt course grade; boxplots show medians, IQRs, and ranges.

(2022–2024), creating a new variable named period. The response variable was the average essay grade, and explanatory variables were sex (female, male) and period. We fitted a linear model (lm in R) including main effects and interaction: average grade ∼ Sex ∗ period. We examined whether sex differences in average essay grades varied across the three periods by evaluating the interaction terms. Additionally, we performed pairwise contrasts using the emmeans package to compare female and male students within each period and obtained p-values adjusted for multiple comparisons.

We analyzed a second dataset containing essay-level data for each student, including student ID, semester, year, course, sex (female, male), essay number (1 to 4), and grade received for each essay. To test whether male students typically submitted fewer essays compared to female students, we counted the total number of essays submitted per student and modeled the count using a Poisson regression with sex as the explanatory variable. We extracted the model-estimated mean number of essays by sex, along with 95% confidence intervals.

2.4. Statistical Tools and Models

To evaluate whether students' grades improved over the course of the semester and whether the rate of improvement differed between sexes, we fitted a linear mixed-effects model using the lmer function from the R package lme4. The model included grade as the response variable, fixed effects for essay number, sex, and their interaction (Sex × Essay Number), and a random intercept for student ID to account for repeated measures: Grade ∼ Sex ∗ Essays + (1 | ID). Model assumptions were checked by visual inspection of residual plots, Q-Q plots,

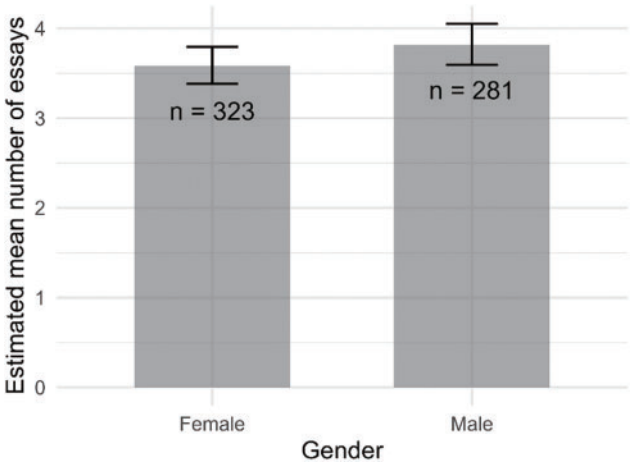

Fig. 2. Estimated mean number of essays submitted by female and male students, with 95% confidence intervals, based on a Poisson regression model.

and histograms of residuals. Although the residual Q-Q plot showed mild deviations from normality, a parametric bootstrap (using bootMer from the lme4 package with 1000 simulations) was performed to obtain robust 95% confidence intervals for all fixed effects.

To test whether the rate of improvement across essays depended on both sexes and a student's initial performance, we categorized students according to the grade they received on their first submitted essay. For each student, we extracted their first essay score and created a binary variable (FirstGradeCat) indicating whether the grade was below 5 ("Low") or 5 and above ("High"). We then fitted a linear mixed-effects model using the lmer function from the lme4 package. The response variable was essay grade, and fixed effects included Gender (Sex), Essays (1–4), FirstGradeCat, and all interaction terms. A random intercept for student ID ((1 | ID)) accounted for repeated measures. The model thus tested whether the effect of essay number on grades varied across both sex and first-grade category: Grade ~ Gender * Essays * FirstGradeCat + (1 | ID). Model assumptions were assessed via residual plots and Q-Q plots. To strengthen inference, we extracted 95% confidence intervals for fixed effects using 1000 parametric bootstrap resamples (bootMer). A likelihood ratio test was conducted to compare the full model (with the three-way interaction) to a reduced model including only two-way interactions. Predicted values were extracted using ggeffects and plotted to visualize performance trajectories by group.

To test whether essay improvement trajectories were influenced by pandemic-related period, sex, and students' initial performance, we fitted a linear mixed-effects model using the lmer function in the lme4 package. The model included fixed effects for Gender, essay number (Essays, 1–4), first essay grade category (FirstGradeCat: low < 5 vs. high ≥ 5), and period (pre-covid, covid, post-covid), as well as all two-, three-, and four-way interactions. A random intercept for student ID accounted for repeated measures. We used the following model: Grade ~ Gender * Essays * FirstGradeCat * period + (1 | ID). We tested model assumptions using residual and Q-Q plots and obtained 95% confidence intervals for fixed effects using 1000 parametric bootstrap simulations (bootMer). To assess whether the four-way interaction significantly improved model fit, we compared the full model to a reduced model (excluding the four-way term) using a likelihood ratio test. Model-predicted grades were visualized across essay numbers by gender (sex), first-grade category, and period using the ggeffects and ggplot2 packages.

3. Results

Across all periods, female students scored higher on average essay grades than male students, with an overall mean difference of approximately 0.8 points (linear model: $p < 0.001$). When examining periods separately, pairwise contrasts revealed that the sex difference was consistently significant. During pre-covid, female students scored on average 0.78 points higher than males ($p = 0.003$); during covid (2020–2021) the difference increased, with females scoring 1.22 points higher ($p = 0.001$). Finally, the difference in grades obtained post-covid was smaller but still significant, with females scoring 0.60 points higher ($p = 0.034$; Fig. 1). These results indicate that the gender (sex) gap in essay performance persisted across all periods, including during the covid lockdown years. Boxplots display the median (center line), interquartile range (box), and range (whiskers). Annotated p-values indicate the significance of differences between female and male students within each period, based on pairwise contrasts from estimated marginal means ($p < 0.01$; $p < 0.05$). Sample sizes (n) per group are shown below each boxplot.

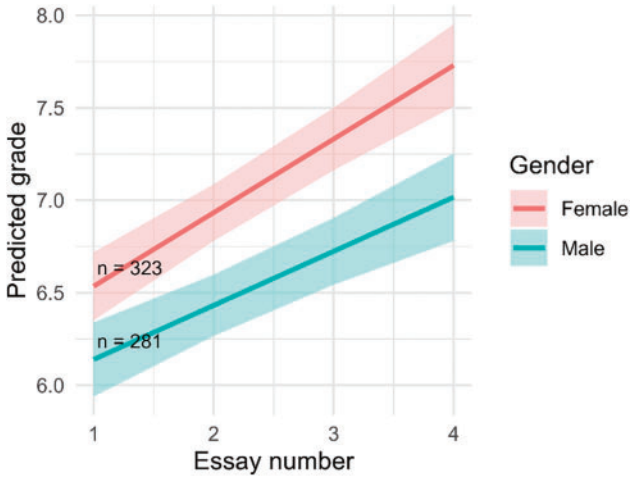

Fig. 3. Model-predicted average essay grades across four consecutive essays, shown separately for female and male students, with 95% confidence intervals.

We tested whether male students submitted fewer essays compared to female students using a Poisson regression model. The model estimated the log-average number of essays for female students as 1.28 (SE = 0.03), corresponding to approximately 3.58 essays (exp(1.28)), and for male students as 1.34 (1.28 + 0.063), corresponding to approximately 3.81 essays (exp(1.34)). However, the difference between sexes was not statistically significant (SexMale coefficient: 0.06 ± 0.04, z = 1.49, p = 0.14), indicating that female and male students submitted, on average, a similar number of essays during the course of ten years (2015–2024) for both classes (Fig. 2). Bars represent the model-predicted average essay count per student; error bars indicate the uncertainty around these estimates. Sample sizes (n) indicate the number of students included in each sex group. No significant difference was detected between sexes in the number of essays submitted.

The linear mixed-effects model showed that grades significantly improved across essays (estimate: 0.40, 95% CI: 0.31–0.49), indicating that students, on average, increased their grades over time. Neither the main effect of sex (estimate: –0.29, 95% CI: –0.66 to 0.08) nor the Sex × Essay Number interaction (estimate: –0.11, 95% CI: –0.24 to 0.03) were statistically significant, suggesting that male and female students improved at similar rates across the semester (Fig. 3). Bootstrap confidence intervals reinforced these findings, confirming the absence of a significant difference in grade improvement slopes between sexes.

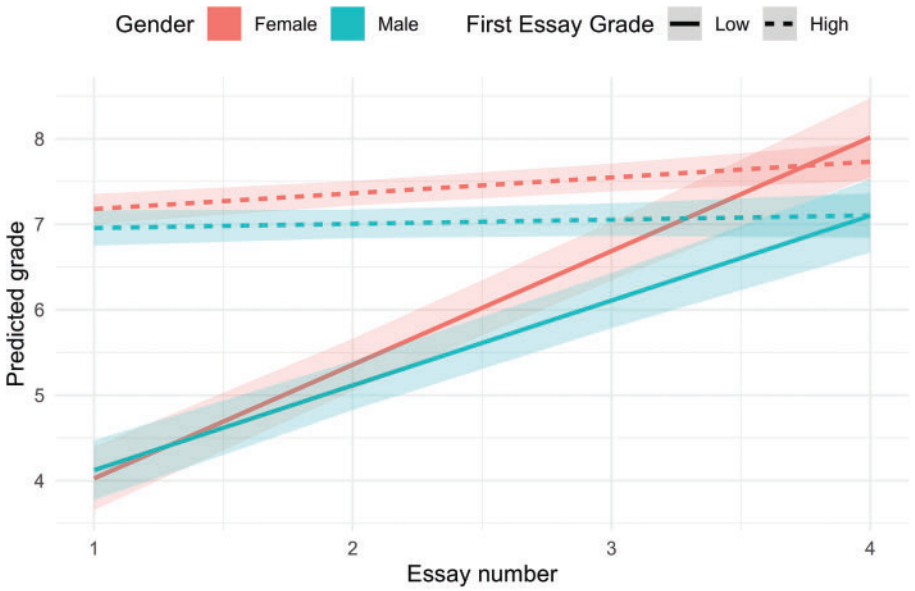

Fig. 4. Predicted essay grades by gender and initial performance level across four essays. Solid lines show trajectories for students starting below grade 5; dashed lines show those starting at or above 5. Shaded areas represent 95% confidence intervals.

Despite the non-significant interaction, model predictions revealed that on the first essay, female students were predicted to score an average of 6.54 (95% CI: 6.35–6.72), while male students were predicted to score 6.14 (95% CI: 5.94–6.34), a difference of approximately 0.4 points in favor of females. By the fourth essay, predicted grades rose to 7.73 (95% CI: 7.51–7.95) for females and 7.02 (95% CI: 6.78–7.25) for males, expanding the grade gap to approximately 0.7 points. These predicted values suggest that although the statistical tests did not detect a significant difference in improvement rates, female students not only began the semester with slightly higher grades but also ended with a larger overall improvement compared to male students. In Fig. 3, lines represent predicted grade trajectories based on a linear mixed-effects model with essay number and sex as predictors, controlling for student ID as a random effect. Sample sizes (n) for each sex are displayed near the starting point of each line (Essay 1). Although both groups show improvement over time, no significant difference was detected between sexes in the rate of grade improvement.

The model to test whether the rate of improvement across essays depended on both sexes and a student's initial performance revealed a significant main effect of essay number, with grades increasing on average by 1.33 points per essay (95% CI: 1.15–1.52, $p < 0.001$). Students with high first-essay grades (≥5) had higher overall performance across the semester than those with low initial grades (estimate = 4.30, 95% CI: 3.75–4.84, $p < 0.001$), but improved significantly less across essays (interaction estimate = –1.15, 95% CI: –1.37 to –0.94, $p < 0.001$). We also found that males improved less than females (Gender × Essays interaction: estimate = –0.34, 95% CI: –0.58 to –0.09, $p = 0.009$). However, the three-way interaction among Gender (Sex), Essay number, and FirstGradeCat was not significant (estimate = 0.20, 95% CI: –0.08 to 0.48, $p = 0.17$), and did not significantly improve model fit (likelihood ratio test: $\chi^2 = 1.86$, df = 1, $p = 0.173$). These results suggest that although females tend to improve more than males, and students with low starting grades improve more than those with high starting grades, the pattern of improvement by sex does not depend on initial performance level (Fig. 4).

When we tested the hypothesis that the relationship between essay improvement, sex, and first-essay performance changed across periods (pre-, during, and post-COVID), we found results consistent with previous findings. Grades improved significantly across successive essays (estimate = 1.29, 95% CI: 1.05–1.54, $p < 0.001$). Students who began with higher first-essay grades (≥5) scored higher overall but improved significantly less than those with low initial grades (Essays × FirstGradeCatHigh: estimate = –1.12, 95% CI: –1.41 to –0.85, $p < 0.001$). Male students improved less than females across the semester (Gender × Essays: estimate = –0.45, 95% CI: –0.81 to –0.12, $p = 0.013$). The four-way interaction among sex, first essay grade, essay number, and period was not significant ($p > 0.17$ for all terms), and a model comparison showed that including this interaction did not improve model fit ($\chi^2 = 1.42$, df = 2, $p = 0.49$). Although one three-way interaction (Gender × FirstGradeCat × period [COVID]) reached marginal significance (estimate = 2.34, 95% CI: 0.15–4.54, $p = 0.039$), this effect was isolated and not supported by broader interaction patterns. These results indicate that while both sex and initial performance shape how students improve over time, these trends remained consistent across pandemic and post pandemic periods (Fig. 5). In this model, solid lines represent predictions from a linear mixed-effects model including all two-, three-, and four-way interactions among Gender (sex), Essay number, FirstGradeCat, and period. Shaded areas indicate 95% confidence intervals based on parametric bootstrapping. Each panel corresponds to a distinct period. Within each panel, line color indicates sex, and line type distinguishes students who began with low vs. high first-essay grades.

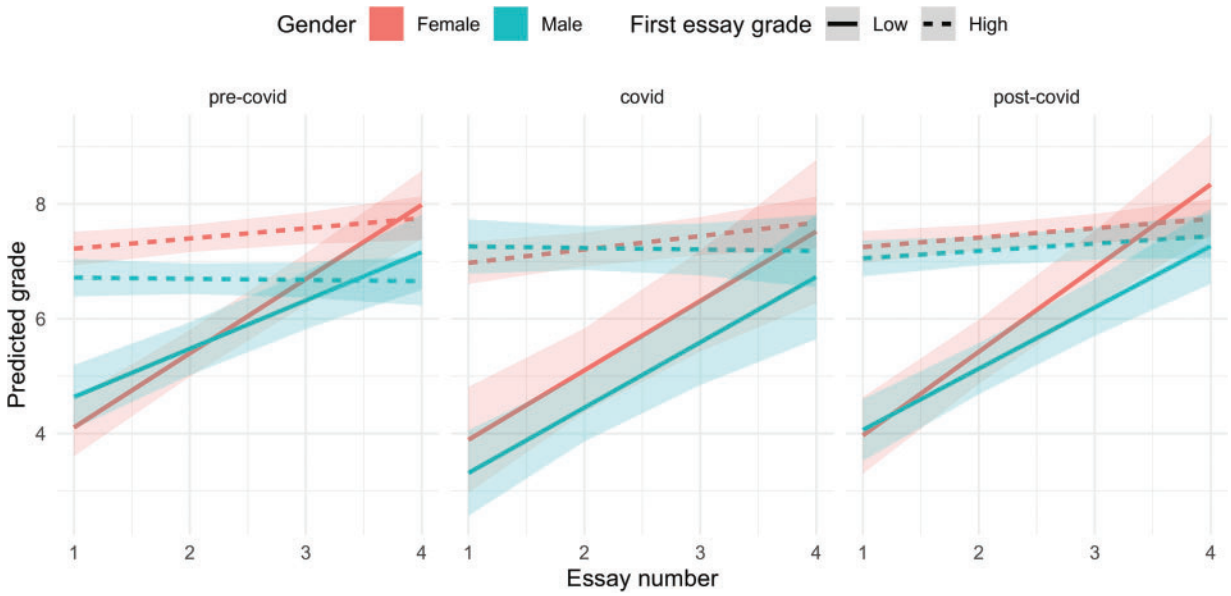

Fig. 5. Model-predicted essay grades across the semester, shown by gender (sex), first essay grade category (low < 5 vs. high ≥ 5), and all three periods.

4. Discussion

There are several issues that stand out from our results. These relate to a persistent sex gap in essay performance, similar improvement trajectories, the impact of the initial first-essay performance, and isolated period-specific effect. All these topics shed some light as they are observed in pre, during and post COVID conditions, including online classes for two years (2020–2021).

The first issue is that female students consistently scored higher than male students in all periods (pre-, during, and post-COVID), with the largest gap observed during the pandemic, as shown in Fig. 1. This is consistent with several studies including Noroozi *et al.* (2023) as they concluded that "female students perform better than male students in argumentative essay writing... and provide higher quality peer review." In our results, the overall difference (∼0.8 points) is statistically significant and suggests a robust trend. However, there is a counter argument: "Male students perform better than female students when taking high-pressure multiple-choice exams, but this gap disappears as the pressure is reduced" (Montolio & Taberner, 2018) [translated]. These findings nuance the role of context and modality in observed sex differences, especially during COVID. Female students are at a disadvantage in the development of mental health issues (Zarowski *et al.*, 2024). This may explain increased strain but not lower performance, which aligns with our finding that women perform better despite stress. Male performance showed a slight upward trend during COVID, but not statistically significantly.

One aspect that may have influenced this difference in results between sexes, according to Espericueta Medina *et al.* (2023), is the fact that, during the pandemic, women had to assume an increased number of domestic responsibilities within their households. Meanwhile men were often faced with the urgency of seeking solutions to financial instability, in many cases working outside the home. Given that virtual platforms became the main mode for university activities during the pandemic, it is possible that women had more consistent access to their coursework, since they remained mostly at home, where internet connectivity tends to be better. Men, on the other hand, may have less access to stable physical space and reliable internet. This is compounded by the fact that, according to Espericueta Medina *et al.* (2023) , men tend to have fewer coping skills when faced with significant life changes. The pandemic introduced many disruptions that may have distracted male students, more than females, from their university duties, potentially contributing to a sharper reduction of their grade improvements.

The second issue is that both sexes improved over time (essay 1 to essay 4). No significant difference in improvement rates across essays when not accounting for initial grade. Whitelaw *et al.* (2024) also noted that performance gains made in 2020 are reversed in 2021... suggesting that 2020 does not reflect a true learning gain. This reinforces our findings that genuine cognitive improvement must be tracked across multiple measurements not assumed from one performance spike. As stated by Noroozi *et al.* (2023) writing argumentative essays requires sustained cognitive processing... performance improves with peer engagement. This also suggests that our findings align with the notion that gradual, essay-by-essay improvements are consistent with deeper learning processes. Such sustained improvement in essay quality, especially in areas of coherence, argumentative structure, and evidence-based reasoning, reflects the progressive strengthening of higher-order cognitive abilities, including analysis, synthesis, and evaluative thinking. These skills are foundational to cognitive development as defined in Bloom's taxonomy and other educational psychology frameworks.

Our findings parallel those of Carson and Kavish (2018), who demonstrated that scaffolding writing assignments across Bloom's taxonomy levels resulted in measurable gains in critical thinking and writing quality within sociology courses. While surface-level improvements in grammar or structure may contribute to some grade changes, the

depth and consistency of the observed gains point toward more substantive cognitive growth. These findings suggest that the improvements observed in student essays are not merely the result of stylistic or mechanical adjustments, but instead indicate measurable growth in metacognitive regulation, argumentation strategies, and abstract reasoning. Such capacities correspond to the upper levels of Bloom's revised taxonomy (e.g., evaluate, create) and illustrate the kind of "knowledge construction" that Anderson *et al*. (2001) identify as central to cognitive development in higher education.

Systematically recording the grades obtained through sustained assessment over a ten-year time-frame allows for the evaluation of whether both the assessment itself and the processes of grading and feedback are producing the expected improvements in student learning. In this case, the data clearly show that essay-based assessment is achieving its intended purpose: students improve as the course progresses. Therefore, tracking this pattern in student grades becomes a valuable resource for instructors to determine whether their course is effectively facilitating the intended learning outcomes, or whether adjustments are needed to enhance student performance throughout the semester.

The robustness of this dataset also strengthens the credibility of the findings. With over 1,800 essays evaluated over a decade using a consistent rubric and a single evaluator, the internal validity of the study is unusually high for educational research. While the data were collected from a single rural campus, the consistency of performance patterns and their convergence with findings from international studies suggest that concerns about generalizability are limited. These results are likely to be applicable in other higher education contexts thereby reinforcing the external relevance of the study.

The third salient issue is the impact of initial performance of the first essay. Students with initial lower grades improved more over time, regardless of sex. Higher starting scores over 5.00 (on a 0–10 scale) were associated with less improvement. This pattern may reflect diminishing returns rather than a true ceiling effect. Students with higher initial grades on the first essay had less room, both statistically and pedagogically, for improvements. Unlike prior studies which focus mainly on outcomes, our essay-by-essay longitudinal model reveals initial grades as a significant determinant of improvement, regardless of sex. As seen in Fig. 4, these results suggest that students with low starting grades improve more than those with high starting grades.

Some motivational explanations may also help to interpret this finding. The more pronounced improvement among students with lower initial scores could stem not only from the greater margin for progress, but also from a sense of urgency or heightened awareness triggered by a poor first performance. This may motivate them to put in extra effort to raise their grades in subsequent assignments. Conversely, students who received higher scores (>5.00) on their first essay may develop a sense of security or overconfidence that discourages further effort toward improvement. It is also possible that teachers, whether consciously or not, provide more targeted feedback and closer follow-up to students with lower initial performance. This is very much in tune with previous studies which demonstrate that formative feedback tend to result in higher student efficacy (Graham *et al*., 2015; Olsen & Hunnes, 2024). The process of given feedback to each essays results in an improved grade in subsequent essays, as seen in our results.

The fourth notable point concerns the interaction between sex and initial performance during the pandemic. When we analyzed COVID period in isolation, the Gender x FirstGrade x Covid-period interaction was marginally significant, suggesting that certain combinations of sex and initial performance, such as low-performing males or high-performing females, might have been more resilient or vulnerable under pandemic conditions. However, this effect was not supported by the full longitudinal model. This marginal difference is only observable when COVID (2020–2021) is tested in isolation. A study by Harris and Reynolds (2023) showed that GPA dropped modestly during COVID for females, but not after. The decline was more pronounced for male students. They were more likely to exhibit performance variability linked to adaptation difficulties or disengagement with online platforms (Alshaibani *et al*., 2023; Kaba *et al*., 2024). This result aligns with our transient COVID-period interaction effects. Similarly, Whitelaw *et al*. (2024) cautions that apparent performance improvements in 2020 may reflect leniency towards grading and reduced content covered on online classes, not actual learning outcomes.

These findings underscore the value of long term studies, highlighting the strength of our longitudinal evidence spanning from 2015 to 2024. They also reinforce the role of sustained, scaffolded essay writing in fostering higher-order thinking skills, an insight particularly relevant for instructors seeking to support cognitive development across diverse learning contexts. Notably, mental health decline was more pronounced for female students during the pandemic, which may have influenced their academic trajectories compared to pre and post COVID (Zarowski *et al*., 2024). Nonetheless, as shown in Fig. 5, our dataset suggest that while both sex and initial performance influenced how students improve over time, these trends remained fairly consistent across pre, during, and post pandemic periods. Overall, the evidence indicates that despite temporary fluctuations during the pandemic (2020–2021), the general trajectory of student improvement remained consistent across the ten-year span.

5. Conclusion

This ten-year longitudinal study offers robust evidence that analytical essay writing supports measurable cognitive development among university students, particularly in arts education. Across both Theatre and Film Appreciation courses, students demonstrated significant improvement in essay performance over time, confirming that sustained written assessment promotes higher-order cognitive abilities such as argumentation, evaluation, and synthesis. As Henry *et al*. (2005) point out, tasks that integrate analytic, creative, and practical skills offer a more comprehensive view of student intelligence and development. In addition to the fact that all essay assignments are

part of art classes with its associated cognitive developments (Li & Qi, 2025), essay writing in particular serve to improve higher-level capabilities. These findings affirm the pedagogical value of essay-based assignments in fostering deep learning, especially within disciplines that rely on critical and interpretive thinking.

Importantly, our results show a persistent sex gap in performance, with female students consistently outperforming male students across pre-, during-, and post-COVID periods. This gap was most pronounced during the pandemic, suggesting possible links to sexed experiences of stress, domestic roles, or access to stable study environments. Nevertheless, both male and female students followed similar trajectories of improvement throughout the semester, particularly when initial performance was taken into account. This is consistent with findings by Noroozi *et al*. (2023), who showed that female students tend to produce more cohesive and well-justified arguments in essay-based assessments.

Students with lower initial essay grades improved more than those who began with high grades, regardless of sex. This effect likely reflects both statistical headroom and motivational dynamics, such as increased effort triggered by a weak early performance or more targeted feedback from the instructor, which is consistent with previous studies (Graham *et al*., 2015; Olsen & Hunnes, 2024). These insights point to the importance of early formative assessment in helping students recognize and act upon their learning gaps.

Although a marginal three-way interaction was found during the COVID period, suggesting that some sex-performance combinations may have been more vulnerable or resilient, the broader longitudinal analysis revealed consistent patterns. These findings support the need for extended multi-year evaluations to capture true learning gains and isolate transient effects, such as those caused by pandemic-related disruptions. In other words, the changes in grade patterns observed during COVID do not represent lasting shifts in cognitive performance, but rather temporary fluctuations probably influenced by contextual and environmental stressors.

Overall, this study highlights the value of long-term, essay-based assessment in monitoring and stimulating cognitive growth and exposing disparities across sex and academic readiness. It contributes new insight from a rural, underrepresented context, where social and economic variables often intersect with educational outcomes. In rural campuses such as UCR at Golfito, educational outcomes are often shaped by structural inequities in access and support (Barquero *et al*., 2021; Vargas *et al*., 2023). While this study draws on data from a rural and historically underrepresented campus, its robust longitudinal design, consistent grading criteria, and alignment with prior international findings strengthen the relevance of its conclusions beyond the immediate setting. Questions of generalizability are rarely raised in studies conducted at elite or urban institutions, despite their limited representation of broader student realities. In contrast, this study provides valuable insight into learning dynamics and equity concerns that are widely applicable across diverse educational contexts (Henry *et al*., 2005; Noroozi *et al*., 2023). On this, we conclude that future research should explore how qualitative factors, such as student motivation, and emotional burden mediate these quantitative patterns. At the same time, future research could strengthen this design by incorporating inter-rater calibration or independent double-coding procedures to assess and confirm the reliability of rubric-based evaluations across multiple scorers. Based on our results, instructors and institutions are encouraged to maintain robust essay assignments and grading processes, and track student trajectories from early coursework to better support cognitive and equitable learning outcomes.

Acknowledgment

We extend our gratitude to Gloriana Chaverri for her invaluable support and assistance with the statistical analysis.

Conflict of Interest

The authors declare that they do not have any conflict of interest.

References

Alshaibani, T., Almarabheh, A., Jaradat, A., & Deifalla, A. (2023). Comparing online and face-to-face performance in scientific courses: a retrospective comparative gender study of year-1 students. *Advances in Medical Education and Practice*, *14*, 1119–1127. https://doi.org/10.2147/AMEP.S408791.

Anderson, L. W. (2005). Objectives, evaluation, and the improvement of education. *Studies in Education Evaluation*, *31*, 102–113.

Anderson, L. W., Krathwohl, D. R., Airasian, P., Cruikshank, K., Mayer, R., Pintrich, P., Raths, J., & Wittrock, M. (2001). A taxonomy for learning, teaching and assessing: a revision of Bloom's taxonomy. In *Artz, AF, & Armour-Thomas, E.(1992). Development of a cognitive-metacognitive framework for protocol analysis of mathematical problem solving in small groups. Cognition and Instruction*, vol. 9, 137–175. Longman Publishing.

Barquero, K., Brenes, V., Lintini, V., León, J., & Murillo, D. (2021). *Estado de la Educación*, pp. 60, Programa Estado de la Nación.

Bertoletti, A., Biagi, F., Di Pietro, G., & Karpiński, Z. (2023). The effect of the COVID-19 disruption on the gender gap in students' performance: A cross-country analysis. *Large-Scale Assessments in Education*, *11*(1), 6. https://doi.org/10.1186/s40536-023-00154-y.

Bloom, B., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of Educational Objectives: The Classification of Educational Goals*. David McKay Company.

Bratti, M., & Lippo, E. (2022). *COVID-19 and the gender gap in university student performance*. (No. 15456). IZA Discussion Papers. https://doi.org/10.2139/ssrn.4178932.

Carson, L., & Kavish, D. (2018). Scaffolding rubrics to improve student writing: preliminary results of using rubrics in a sociology program to enhance learning and mechanical writing skills. *Societies*, *8*(2), 34. https://doi.org/10.3390/soc8020034.

Espericueta Medina, M. N., Sánchez Rivera, L., Villarreal Soto, B. M., Ramos Jaubert, R. I., & López Solís, U. I. S. (2023). La pandemia por COVID-19, su implicación con las emociones y las barreras de aprendizaje en educación superior. *Revista Educación*, *47*(2). https://doi.org/10.15517/revedu.v47i2.52014.

Ferrer, J., Iglesias, E., Blanco-Gutiérrez, I., & Estavillo, J. (2023). Analyzing the impact of COVID-19 on the grades of university education: A case study with economics students. *Social Sciences & Humanities Open*, *7*(1), 100428. https://doi.org/10.1016/j.ssaho.2023.100428.

Graham, S., Hebert, M., & Harris, K. R. (2015). Formative assessment and writing: A meta-analysis. *The Elementary School Journal*, *115*(4), 523–547. https://doi.org/10.1086/681947.

Harris, T. F., & Reynolds, C. L. (2023). COVID-19 diagnoses and university student performance: Evidence from linked administrative health and education data. *Empirical Economics*, *68*(2), 603–637. https://doi.org/10.1007/s00181-024-02653-5.

Henry, P. J., Sternberg, R., & Grigorenko, E. (2005). Capturing successful intelligence through measures of analytic, creative, and practical skills. In *Handbook of understanding and measuring intelligence* (pp. 295–311).

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review*, *2*(2), 130–144. https://doi.org/10.1016/j.edurev.2007.05.002.

Kaba, A., Eletter, S., ElRefae, G. A., Alshehadeh, A. R., & Al-khawaja, H. A. (2024). Students' academic performance before, during, and after COVID-19 in F2F and OL learning: The impact of gender and academic majors. *International Journal of Data and Network Science*, *8*(2), 667–678. https://doi.org/10.5267/j.ijdns.2024.1.011.

Li, J., & Qi, Y. (2025). Arts education and its role in enhancing cognitive development: A quantitative study of critical thinking and creativity in higher education. *Cognitive Development*, *74*, 101544. https://doi.org/10.1016/j.cogdev.2025.101544.

Montolio, D., & Taberner, P. A. (2018). *Diferencias de género en pruebas test bajo presión y su impacto en el rendimiento académico: Un diseño cuasi-experimental*|, *30*, 2, Institut d'Economia de Barcelona.

Noroozi, O., Banihashem, S. K., Taghizadeh Kerman, N., Parvaneh Akhteh Khaneh, M., Babaee, M., Ashrafi, H., & Biemans, H. J. A. (2023). Gender differences in students' argumentative essay writing, peer review performance and uptake in online learning environments. *Interactive Learning Environments*, *31*(10), 6302–6316. https://doi.org/10.1080/10494820.2022.2034887.

Olsen, T., & Hunnes, J. (2024). Improving students' learning—the role of formative feedback: Experiences from a crash course for business students in academic writing. *Assessment & Evaluation in Higher Education*, *49*(2), 129–141. https://doi.org/10.1080/02602938.2023.2187744.

Vargas, J., Román, I., Barquero, K., Lentini, V., León, J., Murillo, D., & Román, M. (2023). *Estado de la Educación*, *9*, 72, Programa Estado de la Nación.

Whitelaw, E., Branson, N., & Leibbrandt, M. (2024). Learning in lockdown: University students' academic performance during COVID-19 closures. *South African Journal of Economics*, *92*(2), 135–160. https://doi.org/10.1111/saje.12367.

Zarowski, B., Giokaris, D., & Green, O. (2024). Effects of the COVID-19 pandemic on university students' mental health: A literature review. *Cureus*, *16*(2), 1–9. https://doi.org/10.7759/cureus.54032.