**Effect of Assessment Structure on Perceived Efficacy in a**

**Rocketry Course**

# Abstract

This study investigates the impact of assessment structure on student performance and engagement in an academic setting, specifically focusing on an introductory rocketry course for undergraduate non-aerospace engineering students. Departing from traditional end-of-course assessments, the research explores whether implementing a 'chunking' approach by breaking the final assessment into individual quizzes over the last week yields distinct outcomes. The approach involved comparing two groups of students: one undergoing a traditional cumulative assessment (Group A) and the other experiencing the modified ‘chunking’ assessment structure (Group B). Paired T-tests were employed to compare the results between the two groups. Both groups had similar demographics and scores prior to the assessment. The results reveal that Group B outperformed Group A with on average a 24% increase in final assessment scores. Additionally, Group B exhibited higher levels of engagement with the material during the assessment week. These findings suggest that modifying the assessment structure by dividing the final assessment into multiple portions may reduce cognitive and testing fatigue, leading to improved student performance and increased engagement.

# Introduction

The pursuit of effective assessment strategies in education is an ongoing challenge. Educators and researchers are constantly on the lookout for methods that not only accurately gauge student knowledge but also promote learning and retention. Traditional end-of-course assessments, long-standing fixtures in academic evaluation, are being re-examined in the context of our evolving understanding of cognitive processes. Ackerman and Kanfer's (2009) research into test length and cognitive fatigue revealed that while students may feel more fatigued the longer they work on tasks, this does not necessarily lead to a decline in performance, suggesting that individual differences and self-regulation strategies play a significant role in educational outcomes. This is further supported by Glaser and Insler's (2022) findings that a day-long break during exams can significantly improve grades, indicating the beneficial effects of rest on cognitive performance. Ackerman et al. (2010) also found that despite subjective feelings of fatigue, performance did not decrease even after extensive testing periods, pointing to the resilience of cognitive function and the influence of individual traits on fatigue and affective responses. These studies collectively underscore the need to rethink traditional assessment methods in favor of strategies that consider the complex dynamics of cognitive fatigue and individual differences.

Chunking is an educational strategy that enhances the learning process by breaking down complex information into smaller, digestible segments, which is especially beneficial in online learning environments. According to the University of Massachusetts Amherst Center for Teaching and Learning, chunking helps students process and understand material more efficiently by dividing content into manageable parts (University of Massachusetts Amherst, n.d.). In the absence of a traditional classroom setting, chunking has been shown to maintain student engagement and attention. Humphries and Clark (2021) found that students preferred chunk-style videos, which led to higher engagement and better completion rates, and even appeared to improve assessment outcomes. Similarly, Faith (2023) reported that chunking course materials improved exam pass rates, outperforming other methods such as expanded homework or automated feedback. These studies suggest that chunking not only facilitates learning but also has a direct positive impact on student success.

The effectiveness of cumulative assessment in educational settings is a nuanced issue, with studies by Domenech et al. (2015) and Kerdijk et al. (2015) highlighting its benefits in improving student performance, motivation, and study habits, while Cecilio-Fernandes et al. (2018) and Tio et al. (2016) present more complex findings. Domenech et al. (2015) advocate for Frequent Cumulative Testing (FCT) due to its positive impact on student engagement and learning outcomes, including higher academic performance and pass rates. Kerdijk et al. (2015) support this by showing that cumulative assessment encourages more self-study time and better performance on recent material, suggesting a potential for improved long-term retention. However, Cecilio-Fernandes et al. (2018) found no significant difference in knowledge acquisition between cumulative and end-of-course assessments, though they acknowledged the benefits of repetitive testing and spaced study sessions. Tio et al. (2016) further complicate the picture by indicating that cumulative assessment may not be as effective for students with weaker self-directed learning skills, pointing to the need for additional support mechanisms within the cumulative assessment framework.

This study investigates the impact of assessment structure on student performance and engagement in an introductory rocketry course designed for undergraduate non-aerospace engineering students. Central to this investigation is whether an assessment approach based on the principles of chunking leads to different and potentially better outcomes. The key questions guiding this research are:

1. Does assessment structure significantly impact outcomes related to cognitive fatigue, considering variations in question timing—either in concentrated sessions or spread out over an extended period?
2. Can incorporating best practices into the assessment structure lead to distinct and potentially improved results?

# Literature Review

## Assessment Strategies

Educational assessment strategies are a key area of interest in contemporary research, particularly in relation to enhancing student performance and engagement. Ghosh et al. (2020) demonstrated that authentic assessments, which simulate practical scenarios, significantly enhance academic achievement, particularly for students with relevant work experience. Fawns and O’Shea (2018) argue for assessments that foster adaptability and critical thinking, essential in our rapidly changing world. Den Boer et al. (2021) compared Summative Cumulative Assessment (SCA)—where grades contribute to final exam scores and emphasize performance for certification—with Formative Cumulative Assessment (FCA), which focuses on feedback for learning improvement without affecting final grades. While SCA showed a short-term improvement in exam grades, its long-term benefits were inconsistent, suggesting that FCA's approach to continuous learning might offer more sustainable academic development. These studies collectively advocate for a shift from traditional to more dynamic and authentic assessment methods to better prepare students for future challenges.

Research also shows that engaging assessment methods can have a positive effect. Georgiou and Nikolaou (2020) couched their assessments within the context of an engaging online game. They observed that these gamified assessments elicited favorable reactions from participants. This preference for engaging methods is mirrored in Sletten's (2021) research. In their study, students found learning more relevant and were more engaged when traditional exams were replaced with paper reviews in a microbiology course. Holmes (2018) found that such methods in virtual learning environments lead to higher engagement and better examination outcomes. Similarly, Andrews et al. (2018) reported improvements in student experience and module grades following an assessment overhaul that included student collaboration in the Portsmouth School of Architecture. These findings suggest that innovative and continuous assessment strategies are crucial for fostering student interaction, confidence, and academic success.

Personalized feedback and alternative assessments methods are known to be received positively by students. The studies by Jayasinghe et al. (2015) and Cirit (2015) explored the effects of personalized feedback and alternative assessment methods. Jayasinghe et al. advocate for monitoring students' emotional states to provide meaningful feedback, proposing discreet observation to tailor educator responses to individual learning experiences and needs. Cirit noted an increase in positive attitudes towards online and alternative assessments among English Language Teaching (ELT) pre-service teachers, suggesting that with thoughtful planning and training, the challenges of implementing new assessment strategies can be navigated successfully. McKevitt (2016) highlights the significant role of specific and timely self-assessment and tutor feedback in enhancing student performance. Similarly, Huisman et al. (2018) underscore the effectiveness of personalized feedback, noting a preference for teacher over peer feedback, and advocate for adaptive feedback systems to boost learning outcomes.

## Cumulative Assessments

There is consensus on the benefits of cumulative assessments for comprehensive learning outcomes. Cumulative assessments, employed across diverse educational settings, evaluate students' overall understanding and retention of knowledge and skills acquired over time. This cumulative assessment approach has benefits for assessing comprehensive learning outcomes. Cumulative assessment generally covers material from multiple courses, facilitates a comprehensive measure of learning progression, and can encourage deeper understanding (Vyas et al., 2015). Cumulative assessments ensure minimal competency, identify knowledge gaps, and foster accountability for cumulative knowledge and skills (Vyas et al., 2015). These assessments play a crucial role in shaping comprehensive learning outcomes and aiding educators in refining instructional strategies (den Boer et al., 2021). Various forms of cumulative assessments, such as exams or projects, prompt students to integrate knowledge, enhancing critical thinking and problem-solving skills. Educators benefit by gaining insights into the effectiveness of teaching strategies and curriculum design (Muniasamy et al., 2015). Implementation of cumulative assessment systems, supported by information technology, enhances academic performance and training quality. These systems monitor students' progress, providing timely feedback to facilitate their learning journey (Kozlov et al., 2019).

## Cognitive Fatigue

Cumulative assessments are often lengthy, and students must spend significant time answering questions and completing tasks. This can lead to cognitive fatigue. Cognitive fatigue, marked by mental weariness, plays a pivotal role in shaping various aspects of performance (Sievertsen et al., 2016). As the day progresses, there is a noticeable decline in student test scores, underscoring the intricate link between cognitive fatigue and academic outcomes (Sievertsen et al., 2016). Importantly, the association between cognitive fatigue, negative well-being, and reduced academic achievement emphasizes the significant and independent influence of cognitive fatigue on performance, even when established predictors are considered (Smith, 2018). The subjective experience of fatigue intensifies with prolonged time-on-task (Ackerman & Kanfer, 2009), highlighting the importance of recognizing time constraints and the availability of cognitive resources when assessing the impact of cognitive fatigue (Borragán et al., 2017). Regardless of cognitive load, subjective cognitive fatigue increases with task duration (Sandry et al., 2014). Furthermore, the relationship between cognitive fatigue, response bias, and brain activation adds complexity to the understanding of cognitive fatigue's impact on decision-making processes (Wylie et al., 2021). However, it should be noted that some research suggests positive benefits associated with longer testing sessions and lengthier cumulative assessments. Some research suggests that incorporating additional exam items is associated with improved scores and enhanced performance, challenging the conventional beliefs about cognitive fatigue (Jensen et al., 2013). Further, there is research that suggests cognitive fatigue may contribute to and enhance the facilitation of procedural motor sequence learning. This implies a more nuanced relationship between cognitive fatigue and skill acquisition (Borragán et al., 2016).

## Chunking of Exams

While there is consensus on the benefits of cumulative assessments for comprehensive learning outcomes, the educational community diverges on the methodology, debating between an all-at-once approach versus a 'chunking' strategy that segments the assessment over time. The concept of chunking has been recognized as a beneficial strategy in both testing and learning contexts, as evidenced by academic research and institutional practices. Drexel University employs chunking as a testing accommodation, allowing students with variable conditions to complete exams in segments within a designated timeframe, thus maintaining exam integrity and fairness (Drexel University). The University of Massachusetts Amherst Center for Teaching and Learning also supports chunking in the learning process, noting that breaking down complex information into smaller parts aligns with the brain's natural processing methods, reducing cognitive overload and enhancing knowledge retention (University of Massachusetts Amherst Center for Teaching and Learning).

Empirical studies further validate the efficacy of chunking. Thalmann, Souza, and Oberauer (2019) demonstrated that chunking in working memory tasks facilitates more efficient information encoding and retrieval by reducing the load on working memory. In the realm of online learning, Humphries and Clark (2021) found that students preferred chunk-style videos to traditional lectures, with the former leading to higher engagement and better learning outcomes. Faith (2023) observed that chunking course materials could improve exam pass rates, suggesting that this technique, when integrated with other teaching strategies, may enhance student success rates. Colver et al. (2021) reported the commonality of chunked exams in large general education classes, which allow students to achieve learning outcomes in stages. Grando (2023) discovered that chunking case studies into videos significantly improved student learning outcomes in microbiology. Lastly, Lees-Murdock et al. (2024) emphasized the importance of manageable learning chunks in supporting student performance in Biomedical Science programs, leading to increased engagement and improved grades.

# Background

We explored the effect of chunking on cumulative exam performance in a hybrid online and hands-on rocketry course. Two groups of students participated in the course. Group A received a single final cumulative assessment (unchunked) while group B received a chunked final cumulative assessment. The course structure is shown in Figure 1 and Table 1. Figure 1 shows the course structure that is common for both groups A and B, while Table 1 illustrates the different cumulative assessment structure between groups. We implemented a structured approach to assess student engagement and learning outcomes in technical content. The number of questions associated with the surveys and technical quiz assessments is given in Table 2.

The course is structured as follows. It begins with an initial pre-content evaluation of student self-efficacy and interest through surveys. Then students engage with online technical content. Each week is a different technical topic and there are four total technical topics (Introduction to Rocket Hardware, Fundamentals of Rocket Design, Modeling Rocket Mechanics, and Flight Analysis). We administer a pre-content technical quiz at the beginning of each week to gauge baseline understanding before students engage with that week’s content. At the end of the week, after engaging with the content, students complete a mid-content technical quiz that is identical to the pre-content quiz. We repeated the same self-efficacy and interest surveys after the online technical content at about the mid-content point. Students then complete the hands-on part of the course that is building, launching, and analyzing data from a model rocket. The final cumulative exam is administered at the end of the course and, up to this point, both groups A and B have experienced the same structure and timing and presentation of content.

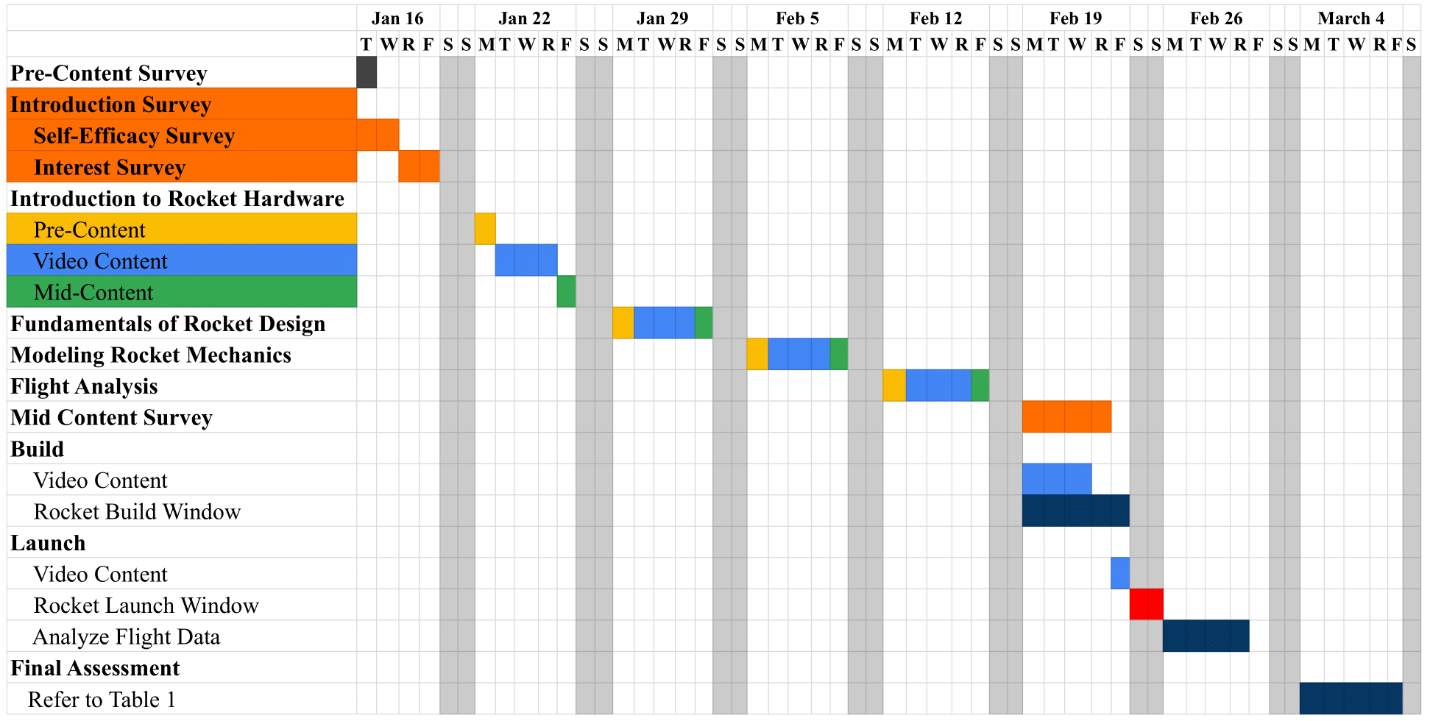


Figure : Common Course and Assessment Structure for Groups A and B

The final cumulative assessment consisted of all course evaluation materials. Specifically, students were asked to again complete the self-efficacy and interest surveys, along with the technical quizzes for all four technical content sections (Table 2). However, the structure and timing of the different surveys and quizzes within this cumulative assessment was different for group A and B. Table 1 illustrates the different cumulative assessment structure between groups. Group A was tasked with a cumulative final exam with a flexible timeline and forced sequential completion of surveys and quizzes. Specifically, students were forced to complete the interest survey, self-efficacy survey, and technical quizzes 1-4 in that order. Note that the lengthy surveys are first. Further, the student was required to complete all these assessments in one sitting because progress could not be saved. The students of Group A had flexibility to complete the cumulative exam task at their discretion any time during the week. Group B was tasked with a "chunked" cumulative assessment. As Table 1 shows, different sections of the cumulative assessment were distributed across the week and required to be completed on specific days. Specifically, the cumulative assessment was divided into three parts: technical quiz 1-2, technical quiz 3-4, and self-efficacy and interest survey. Note that the lengthy surveys are at the end of the cumulative assessment for group B.

Table : Final Cumulative Assessment Structure

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Group** | **Mon** | **Tues** | **Wed** | **Thurs** | **Fri** |
| **A** | Interest, Self-Efficacy Survey, Tech Quizzes 1-4 (in order, at students' discretion) | | | | |
| **B** |  | Tech Quiz 1-2 | Tech Quiz 3-4 | Interest and Self-Efficacy |  |

Table : Assessment Type and Number of Questions

|  |  |  |
| --- | --- | --- |
| **Assessment** | **Number of Questions** | **Type** |
| Self-Efficacy Survey | 39 | 7-point Likert scale |
| Interest Survey | 25 | 5-point Likert scale |
| Introduction and Rocket Hardware | 7 | Multiple Choice |
| Fundamentals of Rocketry | 6 | Multiple Choice |
| Modeling Rocket Mechanics | 6 | Multiple Choice |
| Analysis | 5 | Multiple Choice |

# Logic Model and Theory of Change

Our Logic Model, shown in Table 3, delineates what we believe to be the causal connections between the exam structure and the intended outcomes (student performance and testing fatigue). It categorizes two cohorts: Group A, tasked with an ‘unchunked’ cumulative exam, and Group B, tasked with a chunked exam. Our logic model posits that an unchunked exam correlates with diminished student scores and increased testing fatigue, attributed to inadequate preparation and flexible completion schedule. Conversely, chunked exams yield elevated scores and enhanced readiness, compelling students to complete exams within designated timeframes. The logic model posits that the implementation of chunked exams leads to enhanced student performance and diminished testing fatigue compared to unchunked exams. A restructured exam could lead to better learning outcomes and less fatigue, providing a clearer and more accurate assessment of how students engage with and understand the course content.

Table : Logic Model

|  |  |  |  |
| --- | --- | --- | --- |
| **Group** | **Inputs** | **Outputs** | **Outcomes/Impact** |
| A | * “Unchunked” Final Exam * Students finish at their own pace * All final exam sections due on the same day | * Lower student scores * Less accurate representation of student knowledge | * Superficial understanding of content * Shorter knowledge retention * Less able to use/apply content to new contexts |
| B | * “Chunked” Final Exam * Forced pace to finish on specific days * Final exam sections due on different days | * Students are less stressed and review materials * Higher student scores * More accurate representation of student knowledge | * Deeper understanding of content * Longer term knowledge retention * Able to use/apply content to new contexts |

# Methods

## Researcher Positionality

Our research team is a collaborative effort that includes one master’s student and a professor in aerospace engineering, and one experienced educator with expertise in outreach and educational research. This introductory rocketry course was designed specifically to spark the interest and knowledge of non-aerospace engineering freshmen and sophomores in the world of rocketry and potential space careers. The study presented here delves into a quantitative analysis of the effect of assessment structure on perceived outcomes of participants in this course.

## Procedure

The research was conducted within a spring 2024 course titled "AE298: Introduction to Rocketry" offered by the Aerospace Engineering Department at a major public university in the United States. Recruitment involved various advertising across the university, including departmental emails, strategically placed flyers, and outreach to 4 undergraduate engineering student groups. The target audience was first- and second-year STEM students outside of Aerospace Engineering. The course was graded based on participation. Students received two hours of credit for completing all surveys and quizzes (all participants in this study received full credit). This level of course credit is insufficient to fulfill technical elective requirements, which typically require three or more credits. Self-efficacy questions aimed to gauge students’ confidence in their ability to master the course material and apply their learning. Interest questions, on the other hand, explored their engagement with the topics, personal connections they formed, and intrinsic motivation to delve deeper. This two-pronged approach complemented by qualitative surveys, was designed to glean nuanced insights into student experiences, potentially uncovering hidden gaps or strengths that might elude detection through surveys alone.

## Data Analysis

Survey responses and technical quiz scores were acquired and analyzed. In data analysis, selecting the appropriate statistical test is crucial for accurate conclusions. When data deviates from normality, a rank sum test is preferred, while a parametric t-test is suitable for normally distributed data. Both aim to derive a p-value, with a significance level (alpha) set at 0.01. We first conducted a Shapiro-Wilk test to assess the normality of our data. For normally distributed data, we applied a paired T-test to compare pre- and post-assessment scores within each group, which allowed us to evaluate the progress of individual students. For non-normally distributed data, we utilized the Wilcoxon signed-rank test as a non-parametric alternative to the paired T-test. To compare the two groups' performance, we used an independent T-test (for normally distributed data) or the Mann-Whitney U test (for non-normally distributed data) to determine if the differences are significant. Furthermore, we conducted a power analysis to determine the sample size needed to detect an effect of a given size with a certain degree of confidence. Power analysis is essential to ensure that the study is sensitive enough to detect meaningful differences if they exist; this is particularly important when considering the potential impact of assessment structure on student performance and engagement. By setting our power (1 - beta) at 0.80, we aimed to have an 80% chance of correctly rejecting the null hypothesis when it is false, thereby reducing the risk of Type II errors.

# Results

## Student Demographics

Details and demographics of Groups A and B are shown in Table 4. Group A comprised thirty-two undergraduate students, with 63% being male and 34% female. The majority were either Asian (59%) or white (38%), with 59% being first-year students. Fields of study included mechanical engineering (28%) and physics (25%), the latter being part of the engineering college. Additionally, three students from outside engineering—two from mathematics and one from business—were included. The selection criteria favored early-stage college participants, resulting in 87.5% being freshmen and sophomores, and 12.5% being juniors and seniors.

Group B comprised twenty-six undergraduate students. This group consisted of 58% male and 42% female students, with the majority being either Asian (62%) or white (38%). Furthermore, 58% were first-year students, and the fields of study included mechanical engineering (31%) and astrophysics (15%), with physics being part of the engineering college. Similar to Group A, early-stage college participants were favored in selection, resulting in mostly freshmen and sophomores (85%), with a smaller proportion being juniors and seniors (15%).

Table : Student Demographics

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Group A** | | **Group B** | |
| **Categories** | **n** | **Percent %** | **n** | **Percent %** |
| **Total** | 32 | 100 | 26 | 100 |
| **Gender**  Female  Male  Prefer not to say | 11  20  1 | 34.4  62.5  3.1 | 11  15  0 | 42.3  57.7  0.0 |
| **Ethnicity**  Do not wish to provide  Hispanic or Latino/a  Not Hispanic or Latino/a | 1  6  25 | 3.1  18.8  78.1 | 0  3  23 | 0.0  11.5  88.5 |
| **Race (Multiple selections allowed)**  American Indian or Alaska Native  Asian  Black or African American  White  Do not wish to provide | 1  19  2  12  1 | 3.12  59.9  62.5  37.5  3.1 | 0  10  0  16  0 | 0  38.5  0  61.5  0 |
| **Year in College**  1  2  3  4 | 19  9  3  1 | 59.4  28.1  9.4  3.1 | 15  7  3  1 | 57.7  26.9  11.5  3.9 |
| **Degree Program**  Ag. & Biomed. Eng.  Astrophysics  Astronomy  Civil Eng.  Chem. Eng.  Comp. Eng.  Comp. Sci.  Elec. & Comp. Eng.  Eng. Mechanics  Eng. Undeclared  Ind. & Ent. Sys. Eng.  Mat. Sci. & Eng.  Mech. Eng.  Math  Nuclear Eng.  Physics  Business | 1  1  1  2  1  4  2  9  2  8  1 | 3.1  3.1  3.1  6.2  3.1  12.5  6.2  28.1  6.2  25.0  3.1 | 4  3  2  1  2  1  1  8  1  1  2 | 15.4  11.5  7.7  3.9  7.7  3.9  3.9  30.8  3.9  3.9  7.7 |

## Assessments Prior to the Cumulative Final Assessment

The assessments prior to the cumulative final assessment consisted of the pre-content and mid-content surveys and technical quizzes. Overall, analysis of these assessment results suggests that there are no significant differences in the initial technical knowledge level or growth in knowledge between groups A and B. Further, each group started the course with similarly high levels of interest and self-efficacy, which remained high at the mid-content assessment. Group B consistently reported higher levels of self-efficacy.

### Interest and Self-Efficacy

Figure 2 and Figure 3 together provide a detailed analysis of the changes in interest and self-efficacy for both groups. The hypothesis tests were conducted to compare the change in self-efficacy and interest for both groups during the pre-content and mid-content phase. The null hypothesis for these tests posits that there is no significant change in self-efficacy and interest levels within the groups between pre- and mid-content assessments. The associated p-values from these tests provide the basis for accepting or rejecting this null hypothesis, thereby indicating whether the observed changes in self-efficacy and interest are statistically significant for each group.

In Figure 2, the left plot illustrates a significant decline in interest for Group A from the pre- to mid-content assessment. This result is statistically significant as shown by the data in Figure 3, which quantifies the change in the average score between pre- and mid-content. In contrast, Group B interest levels remain relatively unchanged. The right plot in Figure 2 shows a significant increase in self-efficacy from the pre- to mid-content, which is further supported by the statistical findings in Figure 3. This evidence indicates growth in self-efficacy for both Group A and Group B, suggesting that students' confidence in their abilities to succeed in the course is enhanced as they progress through the material. The combined data from Figure 2 and Figure 3 reveal a nuanced dynamic between interest and self-efficacy within the course content between the two groups. While Group A reports a notable decline in interest, the self-efficacy of both groups significantly increased, suggesting that course content may enhance students' confidence in their academic abilities, even as their interest levels decline.

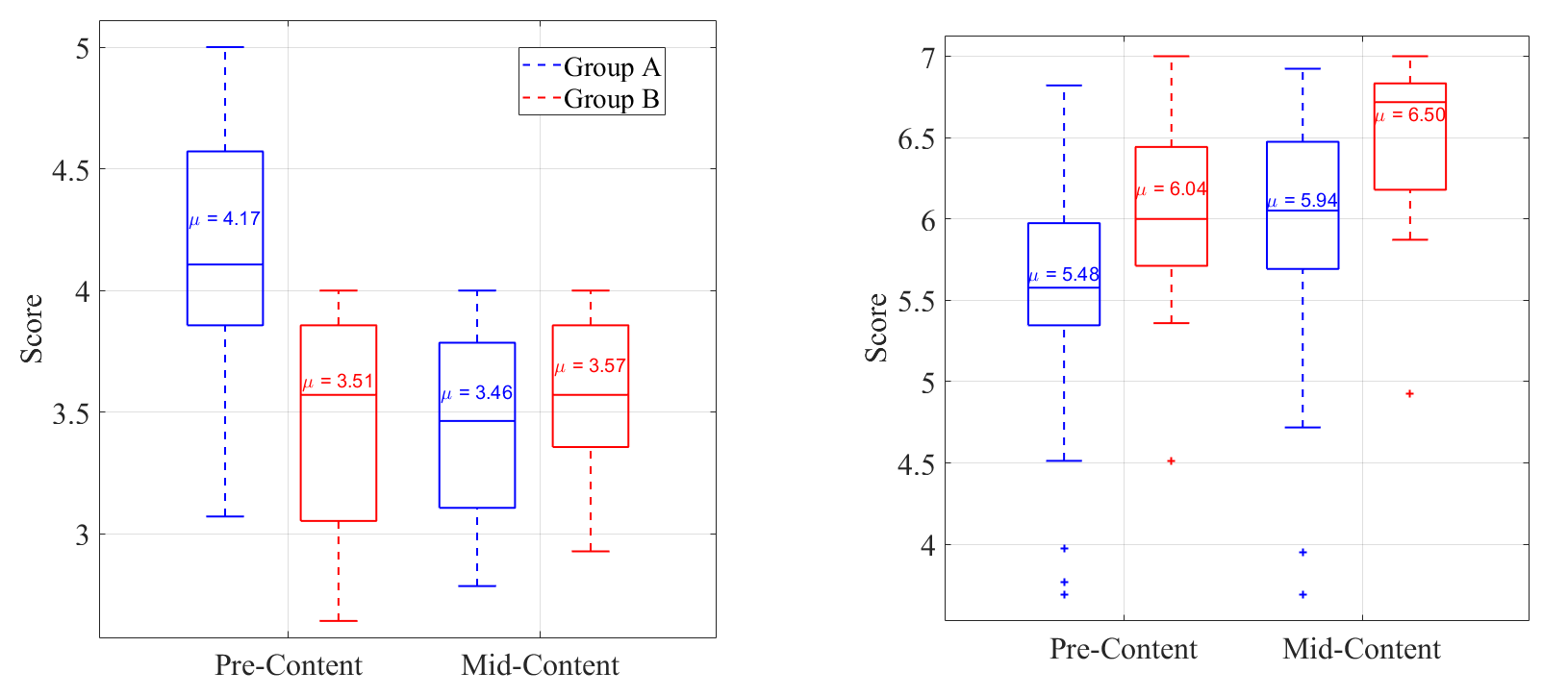


Figure : Interest (left, 5-pt scale) and Self-Efficacy (right, 7-pt scale) prior to the final assessment.

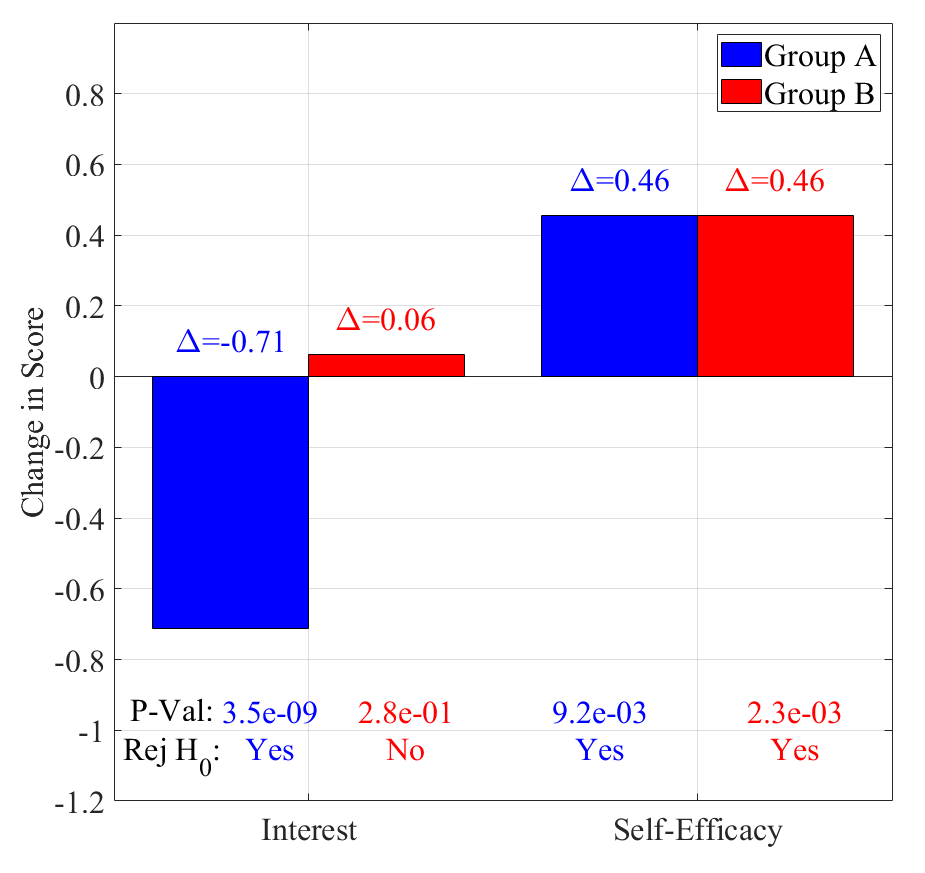


Figure : Changes in Interest and Self-Efficacy from Pre-Content to Mid-Content

### Technical Quizzes

Figure 4 presents a comparison of the distribution of technical quiz scores between the groups for each technical content module, with the mean scores indicated inside each box. During the pre-content phase (left plot in Figure 4), there is a general trend of increasing mean scores across all modules for both groups, except for a decrease in module 3. It is interesting to note that module 3 is more computationally intense, and this may explain the lower mean scores observed. For Group A in module 2, all scores are above the lower quartile, suggesting a higher baseline performance. Group B's scores for modules 2 and 3 are more equally distributed about the median, as indicated by the even spacing between the lower and upper quartiles. Group A's scores for modules 1 and 4 also display this balance, while module 3 shows a wider spread of scores below the median, reflecting greater variability in lower performance.

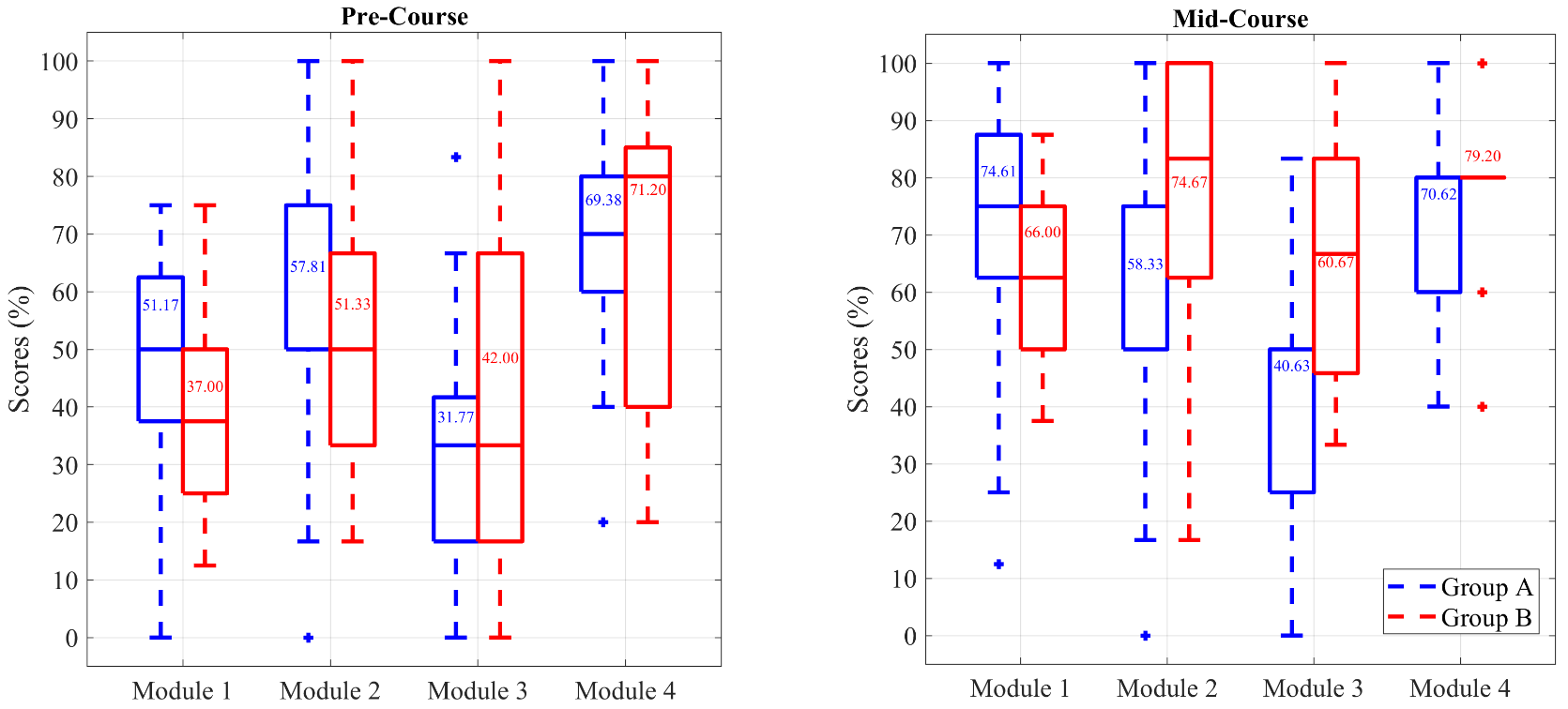
**

Figure : Group A and B Technical Quiz Scores for each Technical Module pre-content (left) and mid-content (right)

In the mid-content phase, as depicted in the right plot of Figure 4, we observe an overall increase in scores and performance for both groups compared to the pre-content assessment phase (left plot). For Group A, module 1 scores are evenly distributed between the lower and upper quartiles. However, for modules 2 through 4, the scores are predominantly clustered towards the upper end of the spectrum. This is indicated by the absence of scores above the upper quartile and, as queried, the median (represented by the middle bar on box and whisker plots) aligns with the upper quartile edge. This phenomenon occurs when the data has certain characteristics, such as outliers or when it is skewed, causing the median to coincide with the quartile boundary. It could imply either a robust grasp of the content or a potential ceiling effect in the assessment's ability to distinguish among top-performing students.

Group B's performance varies, with modules 1 and 3 displaying a lower median than modules 2 and 4. Notably, for module 4, Group B's scores are remarkably uniform, with all scores nestled within the lower and upper quartiles, indicating a consistent performance across the group. Group A presents negative outliers in modules 1 and 2, highlighting areas where a subset of students scored substantially lower than their peers. Conversely, Group B exhibits a broader spectrum of performance in module 4, with one positive outlier and two negative outliers, suggesting a more diverse range of outcomes within that module.

Figure 5 shows the change in mean scores across the modules between the groups, including the p-value and hypothesis decision. For both groups and all modules, the average quiz score increased from pre-content to mid-content assessment. The average quiz performance increased most for module 1, which is the first module and at the beginning of the course. Group B’s performance improvement decreased almost linearly from module 1 to 4 throughout the course. For all modules, group B average quiz score increased more than group A. Group B performance increased on average 19.8% for each module, while group A performance increased only 8.5%.

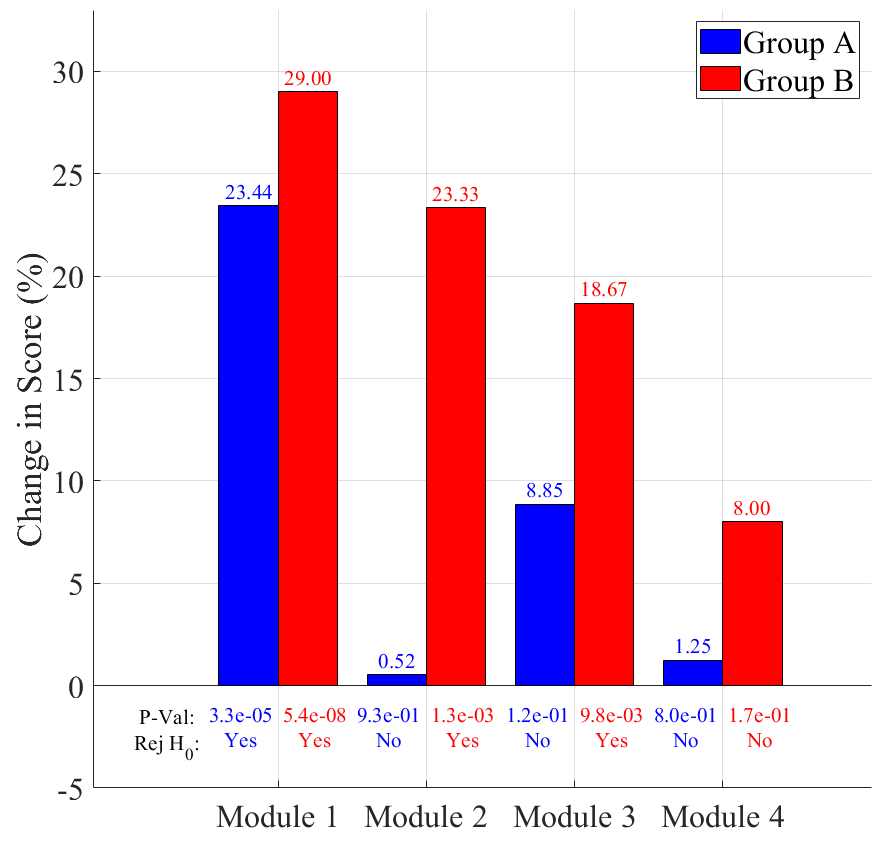


Figure : Change in Average Technical Quiz Score from Pre- to Mid-Content for each of the Four Modules

Scores between group A and B were compared using a two-sided hypothesis test. Table 5 shows these comparisons. The null hypothesis asserts that there are no differences in the mean scores of the two groups across modules and surveys, suggesting that any observed differences are merely due to random variation. The alternative hypothesis contends that there are significant differences between the groups.

Table 5 presents the p-values from our statistical tests alongside decisions on the null hypothesis. For the interest survey in the pre-content phase, the p-value is 4.4e-07, which leads to the rejection of the null hypothesis, confirming that Group A had a significantly higher level of initial interest. In contrast, the mid-content phase showed no significant difference in interest levels between the groups, with a p-value of 2.6e-01. Group B reported significantly higher levels of self-efficacy at both the pre-content and mid-content phases, with p-values of 1.1e-03 and 3.2e-03 respectively. Technical quiz scores showed no significant difference at either phase, supporting the assertion that the groups are similar overall. For all the technical module scores, the p-values do not lead to the rejection of the null hypothesis for both the pre-content and mid-content comparisons. This suggests that, in terms of technical quiz performance, there are not significant differences between group A and B.

Table : Hypothesis Results Between Groups

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Pre-Content** | | **Mid-Content** | |
| **Module/Survey** | **P-value** | **Reject Null** | **P-value** | **Reject Null** |
| **Interest** | 4.4e-07 | Yes | 2.6e-01 | No |
| **Self-Efficacy** | 1.1e-03 | Yes | 3.2e-03 | Yes |
| **Module 1** | 4.6e-01 | No | 7.3e-01 | No |
| **Module 2** | 8.9e-01 | No | 6.7e-01 | No |
| **Module 3** | 8.4e-01 | No | 9.1e-01 | No |
| **Module 4** | 9.4e-01 | No | 1.4e-02 | No |

## Final Assessment Results

Figure 6 left presents results for each module within the technical knowledge portion of the final assessment. For module one, the mean scores for Group A and Group B are similar, at approximately 72% and 71%, respectively. However, Group B out performed Group A on the remaining three modules by on average 13.1%. Figure 6 right illustrates the cumulative final assessment technical module score distributions for Group A and Group B, revealing distinct differences in performance. Group A achieved an average score of 61.6, with a wide range of scores stretching from approximately 35 to 90 percent. The distribution for Group A is skewed towards higher scores, as indicated by an upper quartile that is substantially larger than the lower quartile, suggesting that a significant number of students scored in the higher score range. In contrast, Group B's average score was notably higher at 73.9, and their score distribution was more balanced, with nearly even upper and lower quartiles, which points to a more uniform distribution of scores across the group.

A diagram of a diagram

Description automatically generated with medium confidence

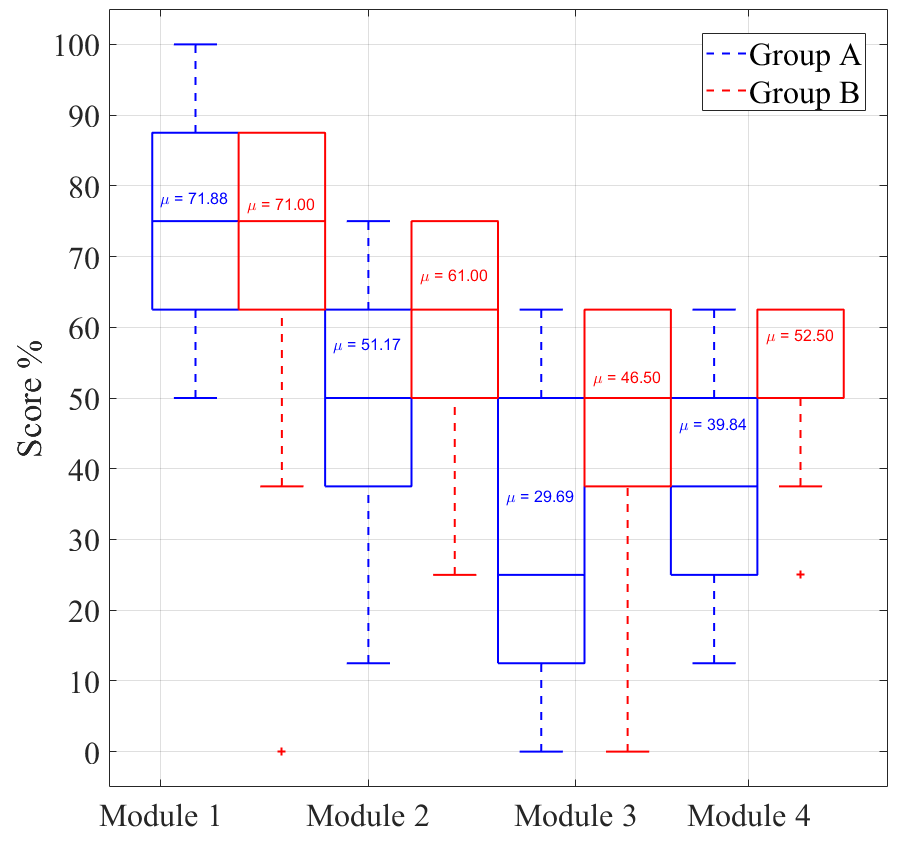


Figure : Final Assessment Technical Quiz Scores by Module (left) and Cumulative (right)

Figure 7 shows the interest and self-efficacy survey results. The same hypothesis testing was employed for this cumulative assessment to determine significant differences between groups and those results are reported in Table 6. Group A reported higher interest, while group B reported higher self-efficacy, and these differences between groups are significant. Focusing specifically on interest surveys, group A reported a broader score range, from 3 to 5, with a mean of 4.17. The upper quartile was larger than the lower quartile, reflecting that more students reported higher levels of interest. Group B's interest scores ranged from 3 to 4, with a lower mean of 3.59, and a larger lower quartile compared to the upper quartile, suggesting that a greater number of students reported lower levels of interest. Group B reported no change in interest throughout the course and reported the same average interest level at the final, 3.59 (Figure 7) as at the mid- and pre-content assessments, 3.57 and 3.51 (Figure 2), respectively. On the final assessment, group A reported a return to their pre-content interest level. Group A’s average final interest level is 4.17 (Figure 7) and is the same as the pre-content level, 4.17 (Figure 2). This is an increase from the average mid-content interest level of 3.46 (Figure 2).

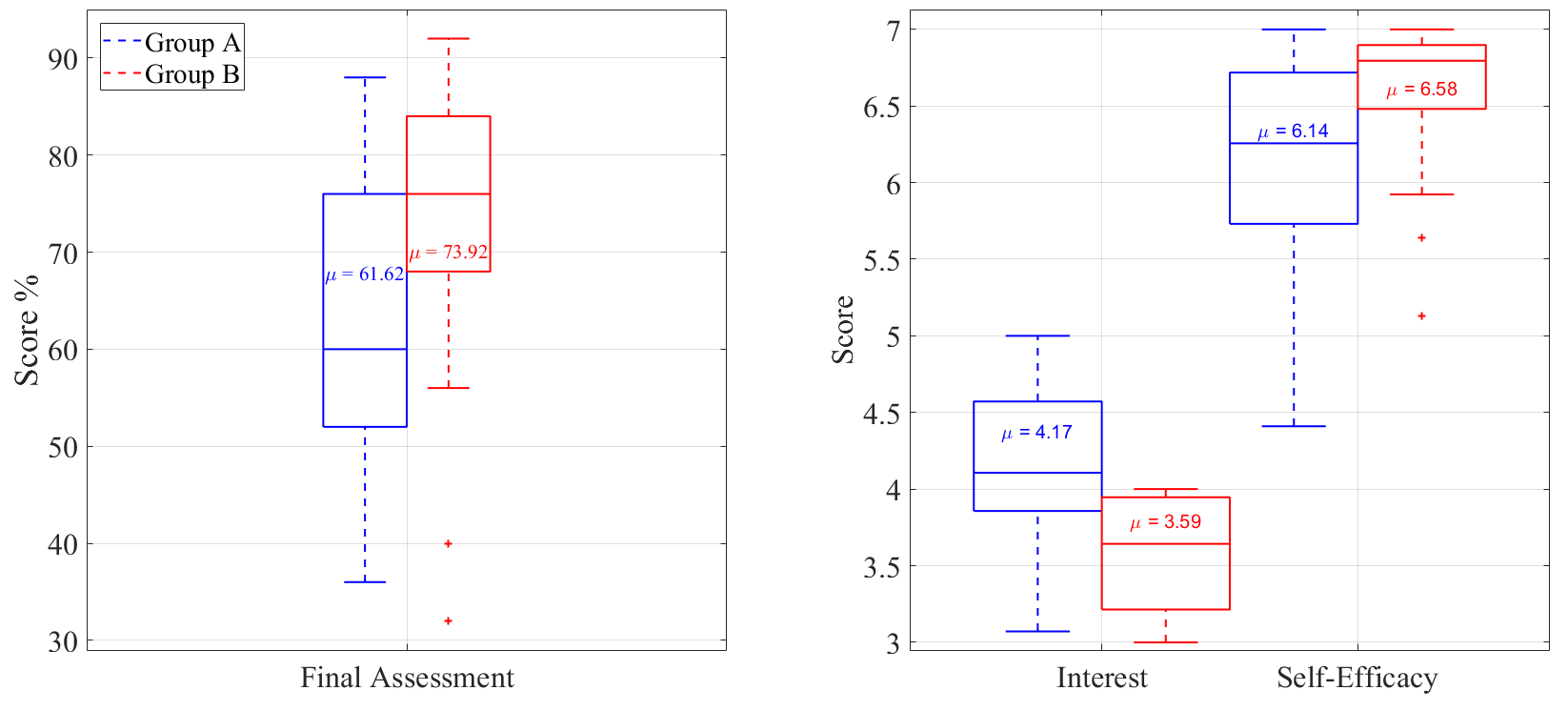


Figure : Final Assessment Results from Interest (5-pt scale) and Self-Efficacy (7-pt scale) Surveys

Table : Final Assessment Score Comparison Hypothesis Results

|  |  |  |
| --- | --- | --- |
| **Module/Survey** | **P-value** | **Reject Null** |
| **Interest** | 1.1e-13 | Yes |
| **Self-Efficacy** | 2.5e-06 | Yes |
| **Technical Quiz Cumulative** | 6.43e-03 | Yes |

## Student Engagement Throughout the Course and Final Assessment

### Online Video Content Access

We investigate student engagement during the course and during the final assessment week by tracking and analyzing course online video access and usage. Specifically, we report here the number of times a video was viewed and the average duration of each viewing. Figure 8 reports these engagement metrics for the four-week period during which the technical modules online video content was assigned (weeks of Jan 22 to Feb 12 in Figure 1). Figure 9 reports the same engagement metrics for the week of the cumulative final assessment (week of Mar 4 in Figure 1).

Figure 8 is a plot of the total views for each course video for each group before the final assessment week. There are a total of 31 videos available sequentially across the four technical modules (M1, M2, M3, M4). A common trend is observed where views are initially high at the start of each module, but tend to decrease as the module progresses. This pattern suggests that students are most engaged at the beginning of a new topic, with interest waning as they become more familiar with the content. However, Module 3 (M3) deviates from this trend, with views increasing towards the end of the module. This uptick in engagement is likely attributed to the last video in M3, which covers a skill directly related to a computationally difficult technical quiz question (related to modeling and plotting a rocket's trajectory). The initial videos of Module 1 show similar view counts to the peak observed in Module 3, indicating a strong start to the course.

Group A, with 32 students, and Group B, with 26 students, started the course with similar engagement, both with over 50 views on the first video. In Modules 1 and 2, Group A consistently had more views, reflecting stronger early engagement, but by Module 3, both groups had similar view counts. At the beginning of Module 4, views were again comparable, but Group B surpassed Group A for the majority of the remaining videos. Despite this shift, by the end of the course, Group A displayed higher interest levels, while Group B showed greater self-efficacy.

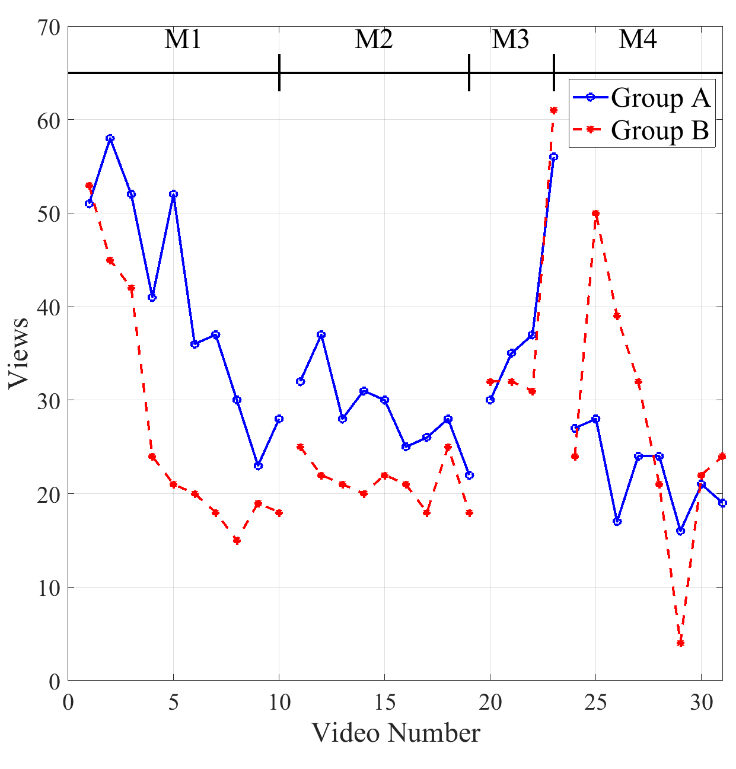


Figure : Number of Views before final assessment

Due to the high number of views for the final video in Module 3 (video 23), we conducted a deeper analysis of its statistics. We found that it had the longest view duration of all the videos, averaging 350 seconds, compared to an average view duration of 200 seconds for the other videos. However, despite having the longest view duration, only 30% of the entire 12-minute video was viewed on average, the smallest fraction of any video. In contrast, the other videos had an average fractional viewing duration of 65%. This is likely due to its utility in addressing two specific quiz questions in Module 3, where students engaged with the video just enough to gather the necessary information, resulting in lower overall view percentage despite its length.

Introductory videos in Modules 2, 3, and 4, which serve as overviews, consistently showed the longest view durations within their modules with an average of 35 and 34 seconds for Group A and B respectively. While the rest of the modules averaged 30 and 25 seconds for Group A and B respectively. This suggests that students may skip these sections in anticipation of more technical content. The first video of Module 1, however, stands out with a notably high view duration, highlighting the significance of early engagement.

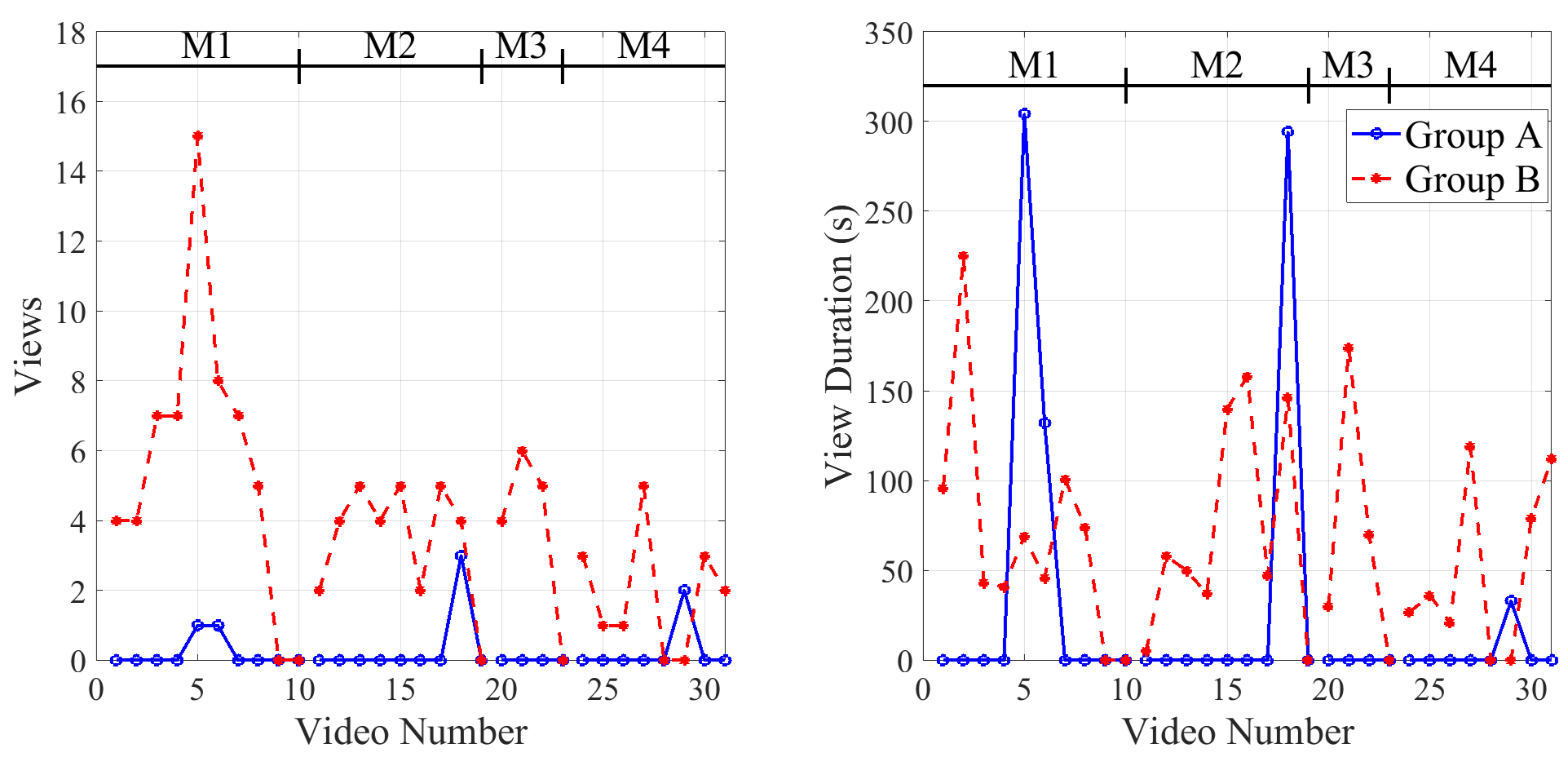


Figure : Views (left)and View Duration (right) During Final Assessment Week

Figure 9 (left) shows video viewing during final assessment week. Group A shows minimal interaction with the course videos, with negligible views across most modules. Interestingly, there are zero views for the video 23 in Module 3, despite its direct relevance to a computationally intensive question on the final. Group B's engagement is higher and more consistent, with multiple views in each module. A notable peak occurs with video 5 of Module 1, which accumulated a total of 15 views. For Modules 2 through 4, Group B's views are generally within the range of 1-6 views, indicating a steady but moderate engagement with the content during finals week. We note that group B also did not engage with video 23.

Figure 9 (right) presents a breakdown of the average view duration per video. The trend observed in Figure 9 left is apparent with Group A showing almost no engagement, resulting in negligible view durations. However, for the few videos that Group A did engage with, they played the entire video, as indicated by the view durations matching the video lengths. This could imply that Group A students were not actively engaging with the content, but rather letting the videos play through to the end. On the other hand, Group B's view durations range from 0 to 200 seconds. None of the view durations for Group B during finals week correspond to the full length of the videos, which may suggest that students are selectively watching specific segments of the videos to review particular topics or find answers for the final exam.

|  |  |  |
| --- | --- | --- |
|  | **P-Value** | **Reject Null** |
| **Views before final** | 8.4e-1 | No |
| **View Duration before final** | 9.1e-1 | No |
| **Views during final** | 9.1e-09 | Yes |
| **View Duration during final** | 1.5e-06 | Yes |

Table : View and View Duration Hypothesis Results

Table 7 presents the results of statistical hypothesis tests conducted to compare the view counts and view durations between the two groups throughout the course and specifically during finals week. For the duration of the course, the null hypothesis posited that there would be no significant difference in the view counts and view durations between the two groups. The alternative hypothesis suggested that a difference did exist. The resulting p-values for the views (8.4e-1) and view duration (9.1e-1) exceeded the predetermined alpha level of 0.01, indicating insufficient evidence to reject the null hypothesis. Consequently, we cannot conclude that there is a statistically significant difference in the engagement metrics between the groups before the final assessment week.

Conversely, during final assessment week, the null hypothesis was that Group B would not exhibit greater engagement with the course content compared to Group A, with the alternative hypothesis stating that there would be no difference in engagement levels. The analysis yielded p-values of 9.1e-09 for views and 1.5e-06 for view duration, both significantly lower than the alpha threshold. This provides strong evidence to reject the null hypothesis in favor of the alternative, suggesting that Group B's engagement with the course content was indeed higher than that of Group A during the final assessment week.

### Final Assessment Submission Metrics

We further investigate student engagement during the cumulative final assessment week by tracking and analyzing submission metrics for the final assessment. Specifically, we report the number of submissions each day, the average technical quiz score associated with submissions on a particular day, and the percentage of the student-exam completed. To clarify, for the 32 Group A students, each individual must complete the entire exam in one sitting. Therefore, one student finishing the exam is equivalent to 1/32 = 3.1% of the total student-exam. In contrast, Group B students are required to complete a section of the exam each day. For instance, if five of the 26 students complete the first of three exam sections, then (5/26)x(1/3) = 6.4% of the entire student-exam has been completed.

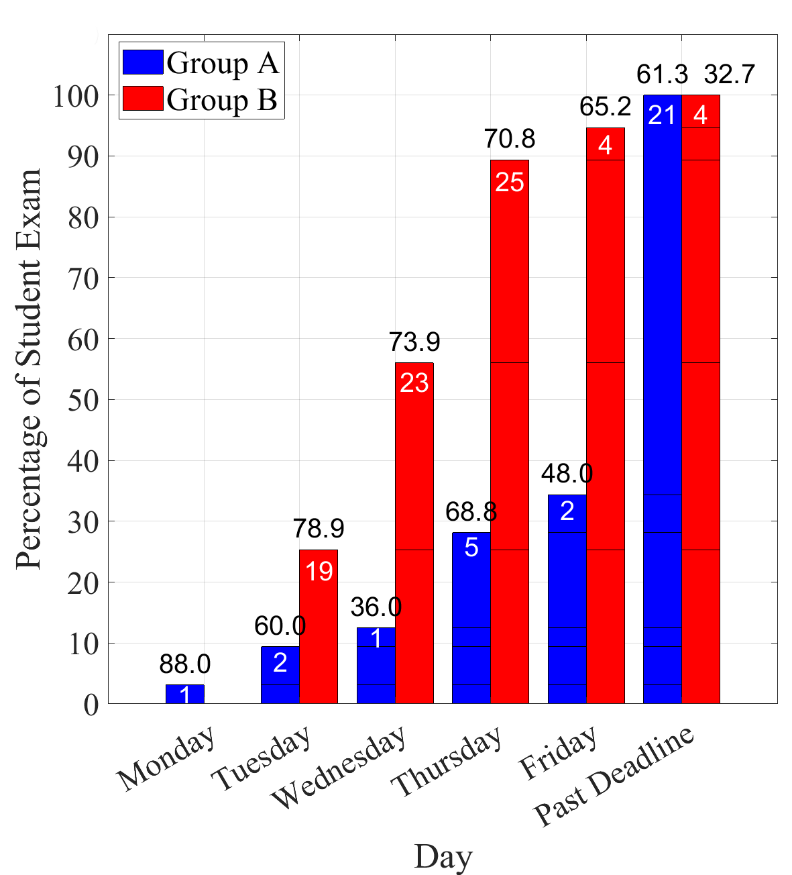


Figure : Final Assessment Week Submission Metrics

Figure 10 shows the submission metrics for the final assessment week. The white number inside the bar is the number of submissions on that particular day. The black number above the bar is the average score for the submissions submitted that day. The height of the bar is the cumulative percentage of student-exams submitted. For example, on Thursday of the final assessment week, there were five group A submissions with an average technical quiz score of 68.8% and these submissions brought the overall student-exam percentage up to 28% (1+2+1+5=9 out of 32). There were 25 group B submissions with an average technical quiz score of 70.8% and these submissions brought the overall student-exam percentage up to 89%. It is important to note that Thursday's session was dedicated to Interest and Self-Efficacy, so no exam scores were recorded. The score presented reflects the average of students who submitted the technical quizzes from Tuesday and Thursday after the deadline.

For Group A, which took the unchunked exam, submission rates fluctuated throughout the week. Initially, only one student completed the exam on Monday, achieving a mean score of 88. By Tuesday, the number of submissions increased to two, but the mean score dropped to 60. Midweek, only one student submitted, scoring a mean of 36. A surge in submissions occurred on Thursday, with five students completing the exam, leading to an improved mean score of 68.8. Two additional submissions were made on Friday, resulting in a mean score of 48. Notably, 21 students submitted their exams after the official deadline, with an average score of 61.333. Group A's completion rate reached 100% after the deadline, indicating a spread-out submission pattern over the week.

In stark contrast, Group B's experience with a chunked exam format, distributed exams and surveys from Tuesday to Thursday, showcased a more uniform pattern of completion. Although the completion rates for Group B fell slightly short of the ideal 33%, 66%, and 100% for Tuesday, Wednesday, and Thursday respectively, they are 25.3%, 56.0%, and 89.3%, respectively. By Friday, the completion rate had marginally risen to 94.7%, with the remaining submissions arriving after the deadline. This structured approach starkly contrasts with Group A's erratic completion timeline and suggests that a chunked exam format may foster more punctual submissions.

Integrating these insights into the previous analysis, we observe that Group B's mean scores began high on Tuesday at nearly 79%, which included only part 1 of the chunked exam (modules 1 and 2). By Wednesday, the mean score decreased to about 74%, reflecting both the second part of the exam and the late submissions from part 1. Thursday's average, devoid of technical questions, further declined to around 70%. This trend indicates that despite the structured exam format, there were still instances of late submissions, which adversely affected the average scores. The 4 students that submitted on Friday submitted their part 3, which includes the Self-Efficacy and Interest Survey, which is why that average isn’t presented on the plot.

The comparison between Group A and Group B is telling. Group A's unchunked format led to a more scattered submission pattern and varied mean scores, while Group B's chunked format seemed to encourage a more consistent completion rate, albeit with a slight decline in mean scores as the week progressed. The structured schedule of Group B, with its more predictable submission pattern, could inform future decisions on exam scheduling and format, aiming to optimize student performance and timely completion.

# Discussion

The primary aim of our data analysis is to meticulously identify and understand any significant disparities in the behavior of the two groups both before and during the final examination period. By closely examining these engagement patterns, we aim to establish a baseline of comparability between the groups. This is a crucial step in ensuring that any observed differences in performance can be confidently attributed to the impact of the modified final assessment structure, rather than pre-existing variations in study behavior or content interaction.

## Interest and Self-Efficacy Survey

We observed that Group A had a broader range of interest scores and a mean of 4.17, suggesting high initial interest with some variability. This aligns with the fluctuating interest levels observed in Group A at the start and end of the course. The moderate correlation between self-efficacy and interest identified by Rottinghaus, Larson, and Borgen (2003) may explain these fluctuations. Their study suggests that self-efficacy can influence the development of interest through mastery experiences, indicating that a moderate level of self-efficacy is necessary to sustain interest. This could imply that Group A's initial engagement was driven by their perceived competence and likelihood of success, as reflected in their self-efficacy mean of 6.14.

Furthermore, the engagement metrics from the course videos (Figures 7 and 8) showed that Group A's interest seemed to wane as the course progressed, which could be a result of their self-efficacy not being sufficiently high to maintain their initial interest. This is particularly evident in their engagement during finals week, where Group A's interaction with the course material was markedly lower than that of Group B, as shown in Figures 9 and 10.

Conversely, Group B's consistently higher levels of self-efficacy, with a mean of 6.58, might have contributed to their persistent engagement and confidence throughout the course. This is supported by their steady engagement with the course content, even during finals week, where they demonstrated a higher level of interaction with the videos. The hypothesis test results from Table 7 further corroborate this, as they indicate a statistically significant difference in engagement between the two groups during finals week, with Group B showing greater engagement.

## Technical Quizzes Scores

In the results section, we reported that the technical quiz scores for Group A and Group B were closely matched, with Group A having a mean score of 81.2% and Group B having a mean score of 82.6% on the pre-content quizzes. The hypothesis tests performed on these scores, as well as the mid-content quizzes, reveal no statistically significant difference between the two groups, with p-values well above the alpha level of 0.01. This outcome indicates that students from both groups began the course with comparable levels of foundational knowledge and continued to progress at a similar rate, as demonstrated by their initial and mid-content quiz scores.

The similarity in scores on the mid-content quizzes, where Group A had a mean score of 76.3% and Group B had a mean score of 77.1%, suggests that both groups have engaged with and comprehended the course material to a similar extent by the midpoint of the course. The absence of a significant disparity in quiz performance from the beginning to the mid-content of the course suggests that the educational content and teaching methods were equally accessible and effective for both groups, allowing students to advance their understanding at a similar pace.

However, as we observed in the final exam completion rates and scores, there was a divergence in performance during finals week. This divergence, particularly the higher engagement and completion rates of Group B as shown in Figures 9 and 10, and their adherence to the chunked exam format, may have contributed to their better performance on the final exam. The lack of significant difference in the quiz scores implies that any variations observed in the final exam scores are likely not a result of initial knowledge gaps or differences in learning during the course but may be due to other factors, such as study habits, time management, or exam preparation strategies that came into play as the course progressed.

## Final Exam Scores

The findings from the final exam hypothesis test, which yielded p-values of 9.1e-09 for finals week views amount and 1.5e-06 for finals week view duration, indicate that Group B's performance surpassed that of Group A. This is consistent with the observed engagement patterns where Group B demonstrated higher interaction with the course content during finals week, as shown in Figures 9 and 10. Furthermore, the hypothesis test for the technical quiz scores, conducted from the onset of the course through to the point after students had interacted with the online content and undergone assessment, showed no significant difference in performance between the two groups. This suggests that students from both groups, who share similar demographic characteristics, maintained equivalent levels of achievement throughout the course.

The key element that seems to have contributed to the higher final exam scores of Group B is the structure of the assessment itself. Given that the technical quiz scores were similar for both groups, the divergence in final exam performance points to the assessment design as a potential factor influencing the outcome. This leads to the inference that the design and format of the final exam were likely influential in creating the performance gap observed between the two groups. It may be that the assessment structure was more aligned with Group B's learning style or study strategies, or that it inadvertently favored the skills or knowledge areas where Group B students were stronger.

This conclusion prompts a closer examination of the assessment methods and their potential impact on student performance. The fact that Group B's engagement with the course content was significantly higher during finals week suggests that the format of the final exam may have been more conducive to their study habits or preparation approach. Understanding how the structure of an assessment can affect outcomes is crucial for ensuring fair and accurate measures of student learning and can guide future improvements in test design to better reflect the abilities of all students.

## Student Course Engagement

The modified assessment structure's impact on student completion and performance on the final exam was quite pronounced. For Group A, the flexibility in choosing when to complete the exam led to a significant number of students postponing their submissions until after the deadline, as detailed in Figure 9. This procrastination was compounded by the exam structure, which required students to complete lengthy Interest and Self-Efficacy Surveys before they could address the technical content. Such a sequence likely contributed to mental fatigue, which, when coupled with the prerequisite of finishing earlier assessments, may have detrimentally affected their performance accuracy—reflected in the lower mean scores on the days leading up to and following the deadline.

Conversely, Group B experienced a more regimented exam schedule with explicit deadlines and a revised format that omitted the initial technical quizzes. This strategic adjustment was designed to minimize fatigue and optimize performance accuracy. By positioning the Interest and Self-Efficacy Surveys at the conclusion of the exam, students were able to concentrate on the technical material without prior cognitive load, thereby potentially enhancing their preparation and overall performance. As substantiated by hypothesis testing, Group B's final exam results were significantly better than those of Group A. This improvement is attributed to the structured exam format, which not only mitigated fatigue through well-defined study segments but also facilitated a more targeted review of the course material. The focused approach allowed students to build a deeper understanding and retention of the topics, which was evidently beneficial to their performance. The comparative analysis of the two groups underscores the efficacy of a structured assessment in promoting better academic outcomes, suggesting a reevaluation of unstructured exam formats in favor of more guided, segmented approaches to learning and evaluation.

## Finals Week Completion Rate

The modified assessment structure's impact on student completion and performance on the final exam was quite pronounced. For Group A, the flexibility in choosing when to complete the exam led to a significant number of students postponing their submissions until after the deadline, as detailed in Figure 9. This procrastination was compounded by the exam structure, which required students to complete lengthy Interest and Self-Efficacy Surveys before they could address the technical content. Such a sequence likely contributed to mental fatigue, which, when coupled with the prerequisite of finishing earlier assessments, may have detrimentally affected their performance accuracy—reflected in the lower mean scores on the days leading up to and following the deadline.

Conversely, Group B experienced a more regimented exam schedule with explicit deadlines and a revised format that omitted the initial technical quizzes. This strategic adjustment was designed to minimize fatigue and optimize performance accuracy. By positioning the Interest and Self-Efficacy Surveys at the conclusion of the exam, students were able to concentrate on the technical material without prior cognitive load, thereby potentially enhancing their preparation and overall performance. As substantiated by hypothesis testing, Group B's final exam results were significantly better than those of Group A. This improvement is attributed to the structured exam format, which not only mitigated fatigue through well-defined study segments but also facilitated a more targeted review of the course material. The focused approach allowed students to build a deeper understanding and retention of the topics, which was evidently beneficial to their performance. The comparative analysis of the two groups underscores the efficacy of a structured assessment in promoting better academic outcomes, suggesting a reevaluation of unstructured exam formats in favor of more guided, segmented approaches to learning and evaluation.

# Conclusion

In conclusion, our comprehensive analysis has highlighted critical insights into the academic performance and engagement levels of two distinct study groups within the course framework. Notably, the final exam scores and completion rates presented a stark contrast between the groups, with Group B outperforming Group A significantly. This difference in performance can be largely attributed to the assessment structure, which played a pivotal role in influencing study behaviors and exam preparedness.

The Interest and Self-Efficacy Surveys revealed nuanced motivational factors that potentially impacted the students' engagement trajectories. Group A exhibited high initial interest that fluctuated, possibly due to varying levels of self-efficacy. Conversely, Group B maintained a consistent sense of self-efficacy and engagement, which likely contributed to their sustained academic performance.

Despite similar technical quiz scores at the outset, indicating an equal footing in terms of knowledge and engagement, the two groups diverged in their final exam outcomes. This divergence underscores the significance of the assessment structure. Group A's flexible approach to exam scheduling and the sequential nature of their exam led to procrastination and possible cognitive fatigue. In contrast, Group B's structured assessment format, which allowed for focused and strategic preparation, proved to be more conducive to their learning style and ultimately resulted in higher performance.

These observations underscore the profound impact that assessment design can have on student outcomes. They highlight the importance of constructing assessments that not only evaluate student knowledge but also align with their study habits and preparation strategies to maximize performance and engagement.

Moving forward, our research raises several pertinent questions that warrant further investigation:

1. How does the performance on initial assessments influence students' engagement with the course material and their study patterns?

2. What is the relationship between early assessment performance and the likelihood of students revisiting the material for additional study?

3. Can regular feedback on performance and proactive intervention strategies, like reminders and encouragement, enhance student engagement and improve outcomes?

4. What are the effects of different assessment formats on a diverse range of learning styles and student demographics?

Addressing these questions could lead to a deeper understanding of educational assessment strategies and their impact on student learning. Through continued research in this area, we can refine our assessment methods to better cater to the diverse needs of students, thereby enhancing the overall educational experience and academic success.

**References**

1. Plomin T, Kim J. The Impact of a Blended Course about Rocketry on Student Knowledge, Self-Efficacy, and Career Interest

2. Ackerman, P. L., & Kanfer, R. (2009). Test length and cognitive fatigue: an empirical examination of effects on performance and test-taker reactions. Journal of Experimental Psychology: Applied, 15(2), 163.

3. Glaser, D. J., & Insler, M. A. (2022). The deleterious effects of fatigue on final exam performance. Economics of Education Review, 90, 102305.

4. Ackerman, P. L., Kanfer, R., Shapiro, S. W., Newton, S., & Beier, M. E. (2010). Cognitive fatigue during testing: An examination of trait, time-on-task, and strategy influences. Human performance, 23(5), 381-402.

5. Humphries, B., & Clark, D. (2021). An examination of student preference for traditional didactic or chunking teaching strategies in an online learning environment. Research in Learning Technology, 29.

6. University of Massachusetts Amherst. (n.d.). How do I chunk content to increase learning? [University of Massachusetts Amherst Center for Teaching and Learning]. Retrieved from https://www.umass.edu/ctl/resources/how-do-i/how-do-i-chunk-content-increase-learning

7. Faith, T. (2023). Assessing Selected Teaching Techniques and Their Impact on Student Success in the Classroom. Teaching and Learning Excellence through Scholarship, 3(1).

8. Cecilio-Fernandes D, Nagtegaal M, Noordzij G, Tio RA. Cumulative assessment: Does it improve students’ knowledge acquisition and retention? Sci Med. 2018;28(4):ID31880. DOI: 10.15448/1980-6108.2018.4.31880

9. Kerdijk, W., Cohen‐Schotanus, J., Mulder, B. F., Muntinghe, F. L., & Tio, R. A. (2015). Cumulative versus end‐of‐course assessment: effects on self‐study time and test performance. Medical education, 49(7), 709-716.

10. Domenech, J., Blazquez, D., De la Poza, E., & Mun͂oz-Miquel, A. (2015). Exploring the impact of cumulative testing on academic performance of undergraduate students in Spain. Educational Assessment, Evaluation and Accountability, 27, 153-169.

11. Tio, R. A., Stegmann, M. E., Koerts, J., van Os, T. W., & Cohen-Schotanus, J. (2016). Weak self-directed learning skills hamper performance in cumulative assessment. Medical Teacher, 38(4), 421-423.

12. Ghosh, S., Brooks, B., Ranmuthugala, D., & Bowles, M. (2020). Authentic versus traditional assessment: an empirical study investigating the difference in seafarer students' academic achievement. The Journal of Navigation, 73(4), 797-812.

13. Fawns, T., & O’Shea, C. (2018, May). Distributed learning and isolated testing: tensions in traditional assessment practices. In Proceedings of the 2018 Networked Learning Conference (pp. 132-139).

14. den Boer, A. W., Verkoeijen, P. P., & Heijltjes, A. E. (2021). Comparing formative and summative cumulative assessment: Two field experiments in an applied university engineering course. Psychology Learning & Teaching, 20(1), 128-143.

15. Georgiou, K., & Nikolaou, I. (2020). Are applicants in favor of traditional or gamified assessment methods? Exploring applicant reactions towards a gamified selection method. Computers in Human Behavior, 109, 106356.

16. Sletten, S. R. (2021). Rethinking assessment: Replacing traditional exams with paper reviews. Journal of Microbiology & Biology Education, 22(2), 10-1128.

17. Singh, K., Bharatha, A., Sa, B., Adams, O. P., & Majumder, M. A. A. (2019). Teaching anatomy using an active and engaging learning strategy. BMC medical education, 19, 1-8.

18. Holmes, N. (2018). Engaging with assessment: Increasing student engagement through continuous assessment. Active Learning in Higher Education, 19(1), 23-34.

19. Andrews, M., Brown, R., & Mesher, L. (2018). Engaging students with assessment and feedback: improving assessment for learning with students as partners. Practitioner Research in Higher Education, 11(1), 32-46.

20. Jayasinghe, U., Dharmaratne, A., & Atukorale, A. (2015, February). Students' performance evaluation in online education system Vs traditional education system. In Proceedings of 2015 12th International Conference on Remote Engineering and Virtual Instrumentation (REV) (pp. 131-135). IEEE.

21. Cirit, N. C. (2015). Assessing ELT Pre-Service Teachers via Web 2.0 Tools: Perceptions toward Traditional, Online and Alternative Assessment. Turkish Online Journal of Educational Technology-TOJET, 14(3), 9-19.

22. McKevitt, C. T. (2016). Engaging students with self-assessment and tutor feedback to improve performance and support assessment capacity. Journal of University Teaching & Learning Practice, 13(1), 2.

23. Huisman, B., Saab, N., Van Driel, J., & Van Den Broek. (2018). Peer feedback on academic writing: undergraduate students’ peer feedback role, peer feedback perceptions and essay performance. Assessment & Evaluation in Higher Education, 43(6), 955-968.

24. Vyas, D., Halilovic, J., Kim, M. K., Ravnan, M. C., Rogan, E. L., & Galal, S. M. (2015). Use of cumulative assessments in US schools and colleges of pharmacy. Pharmacy, 3(2), 27-38.

25. Muniasamy, V., Ejalani, I. M., & Anandhavalli, M. (2015). Predicting the Students Learning Outcome Based on Comparing the Assessment Methods in Diploma E-Commerce Course, Community College, King Khalid University, Ksa. Assessment, 7(7.18), 1-595.

26. Kozlov, V., Alontseva, E., & Guryanov, A. (2019, May). INFORMATION AND EDUCATIONAL ENVIRONMENT FOR MONITORING AND FORMING A CUMULATIVE ASSESSMENT OF A STUDENT’S ACADEMIC WORK IN A SEMESTER. In SOCIETY. INTEGRATION. EDUCATION. Proceedings of the International Scientific Conference (Vol. 5, pp. 383-392)

27. Sievertsen, H. H., Gino, F., & Piovesan, M. (2016). Cognitive fatigue influences students’ performance on standardized tests. Proceedings of the National Academy of Sciences, 113(10), 2621-2624.

28. Rashtchi, M., Abbaszadeh, M., & Tabarraee, N. (2015). Reading Comprehension Testing Techniques and Cognitive Fatigue: A Study on Construct Irrelevant Factors. Br. J. Educ. Soc. Behav. Sci, 6, 196-209.

29. Borragán, G., Slama, H., Bartolomei, M., & Peigneux, P. (2017). Cognitive fatigue: A time-based resource-sharing account. Cortex, 89, 71-84.

30. Borragán, G., Slama, H., Destrebecqz, A., & Peigneux, P. (2016). Cognitive fatigue facilitates procedural sequence learning. Frontiers in human neuroscience, 10, 86.

31. Smith, A. P. (2018). Cognitive fatigue and the well-being and academic attainment of university students. Journal of Education, Society and Behavioral Science.

32. Beck, D. (2023). Virtual school students and state testing: A story of fatigue, irrelevant thoughts and anxiety. Available at SSRN 4665005.

33. Sandry, J., Genova, H. M., Dobryakova, E., DeLuca, J., & Wylie, G. (2014). Subjective cognitive fatigue in multiple sclerosis depends on task length. Frontiers in neurology, 5, 214.

34. Wylie, G. R., Yao, B., Sandry, J., & DeLuca, J. (2021). Cognitive Fatigue and Brain Activation: A Study of the Relationship Between Cognitive Fatigue and Response Bias. Frontiers in Psychology, 11, Article 579188.

35. Jensen, J. L., Berry, D. A., & Kummer, T. A. (2013). Investigating the effects of exam length on performance and cognitive fatigue. PloS one, 8(8), e70270.

36. Rottinghaus, P. J., Larson, L. M., & Borgen, F. H. (2003). The relation of self-efficacy and interests: A meta-analysis of 60 samples. Journal of Vocational behavior, 62(2), 221-236.

37. Niemivirta, M., & Tapola, A. (2007). Self-efficacy, interest, and task performance: Within-task changes, mutual relationships, and predictive effects. Zeitschrift für Pädagogische Psychologie, 21(3/4), 241-250.

38. Lee, W., Lee, M. J., & Bong, M. (2014). Testing interest and self-efficacy as predictors of academic self-regulation and achievement. Contemporary educational psychology, 39(2), 86-99.

39. Chunking of Exams. [Drexel University]. Retrieved from https://drexel.edu/disability-resources/~/media/Drexel/Treasurer-Group/Disability-Resources/Documents/Chunking-of-Exams.ashx

40. Thalmann, M., Souza, A. S., & Oberauer, K. (2019). How does chunking help working memory?. Journal of Experimental Psychology: Learning, Memory, and Cognition, 45(1),

41. Colver, M. C., Thurston, T. N., Wadsworth-Anderson, E., Grewe, J. R., Kohler, D., & Lewis, M. (2021). Improving University Student Course Performance Using Reflective Exam Analysis: An Approach to Developing Academic Proficiency.

42. Grando, D. (2023). Chunking Microbiology Learning Materials Today. (3), 127–130.

43. Lees-Murdock, D., Irwin, R., Graham, J., Hinch, V., O’Hagan, B., & McClean, S. (2024). Assessing the Efficacy of Active Learning to Support Student Performance Across Undergraduate Programmes in Biomedical Science., 81, Article 12148.