The Impact of Time of Day on Engagement and Performance for a University Low-Stakes Assessment

Authors:

Jonathan M. Henriques, M.A.
James Madison University

Jack Gilmore, M.A.
James Madison University

Brian C. Leventhal, Ph.D.
James Madison University

Yu Bao, Ph.D.
James Madison University

ABSTRACT

Student engagement on low-stakes assessments raises concerns about the validity of conclusions derived from those tests. With the growing prevalence of online administration at universities, students often have the flexibility to choose their testing time. This study explores how the self-selected testing time of day influences university students' test-taking engagement and performance in low-stakes assessments. A total of 1,397 university students were categorized into four time-of-day categories: morning, afternoon, evening, and nighttime. The number of rapid guesses was compared to assess engagement. Performance scores were also analyzed for comparison. The findings revealed that students who chose to test in the morning and afternoon exhibited greater engagement and better performance than those who tested in the evening and nighttime. These results underscore the importance of considering testing time when interpreting assessment outcomes, with implications for the scheduling of online assessments to enhance student engagement and performance.

Correspondence E-mail: Henri2jm@jmu.edu

Keywords: low-stakes assessment, test-taking engagement, test-taking effort, response time effort, rapid guessing

When interpreting assessment scores, it is critical to consider whether examinees appropriately engaged throughout the test administration (American Educational Research Association et al., 2014). Assuming high levels of examinee engagement is typically justified during high-stakes tests, where examinees receive benefits or face consequences based on their performance. However, for assessments with low stakes where performance results in no personal benefit or consequence, low test-taking engagement becomes a significant concern. When an examinee does not fully engage during a test to assure their highest possible score, the validity of our test score interpretations is threatened (Wise & DeMars, 2005). For example, higher education assessments conducted to evaluate program-level or institutional outcomes are commonly low-stakes. Therefore, student motivation to engage in the test tends to be low, consequently reducing their test-taking engagement and performance. In fact, disengaged students perform an average of .59 standard deviations lower than engaged students (Wise & DeMars, 2005).

Understanding the factors that drive an examinee's test-taking engagement is essential for identifying strategies to enhance it. One influential framework for explaining the variability in test-taking engagement is Expectancy-Value Theory (EVT). EVT posits that an examinee's expectations for success and their perceived task value significantly influence their effort (Eccles et al., 1983; Eccles & Wigfield, 2002). Expectations for success represent an individual's self-driven beliefs about likely task performance, while perceived task value reflects the importance they assign to succeeding or failing in a task (Eccles & Wigfield, 2002).

An individual's perceived task value is shaped by four key components: attainment value, intrinsic value, utility value, and cost. Attainment value pertains to the personal significance assigned to the task, intrinsic value relates to the enjoyment derived from task performance, utility value considers how well the task aligns with individual goals, and cost encompasses the negative aspects (such as time and effort) associated with task engagement (Eccles & Wigfield, 2002). When the costs of test-taking outweigh the expected benefits, disengagement often ensues, a pattern especially evident in low-stakes testing contexts, where the lack of meaningful consequences leads to consistently lower engagement compared to high-stakes assessments. Even when examinees derive little enjoyment from high-stakes tests, the potential outcomes often incentivize effort, aligning with EVT (Wise & DeMars, 2005). To explore this dynamic, Wise and DeMars (2005), relied on self-report data to assess examinees' perceived importance and effort, offering insight into the motivational process, but raising questions about the adequacy of self-report as a sole indicator of engagement.

While self-report data can provide valuable insights into overall effort indirectly from the test-takers, it assumes that examinees respond truthfully, and it only captures effort cross-sectionally at the end of the test (Wise & Kong, 2005). Examinees might also provide inaccurate information due to social desirability bias, confusion between poor performance and low effort, or low engagement with the self-report measure itself. Additionally, variation in effort during an exam (Pastor et al., 2019)—such as differential effort across different sections—may not be fully captured by self-report measures administered only at the end of the assessment.

To evaluate effort throughout a test, researchers have often relied on item-level behavioral indicators of disengagement. While effort refers to the mental resources an examinee expends toward success, engagement represents the behaviors that indicate such investment (Gorgun & Bulut, 2023). Two common behavioral indicators are solution behavior and rapid guessing behavior. Examinees exhibit solution behavior when they intentionally attempt to determine the correct answer (Schnipke, 1995; Wise & DeMars, 2006), while rapid guessing behavior occurs when examinees answer an item more quickly than it would reasonably take to read, process, and conclude an answer (Wise & Kong, 2005). During high-stakes assessment, rapid guessing often arises when examinees run out of time and need to complete the test quickly (Wise, 2020). In contrast, during low-stakes assessments rapid guessing tends to appear more sporadically throughout the test and signals disengaged test-taking (Wise & Kong, 2005).

Determining Rapid Guessing Behavior

Rapid guesses can be identified by comparing the response time with a predetermined time threshold for each item. Responses faster than this threshold are labeled as rapid guesses. Several methods exist to establish rapid guess thresholds. The simplest approach involves selecting a fixed threshold time and applying it uniformly across all items (Wise et al., 2010). While easy to implement, this method overlooks variations among items as some may naturally require more or less time for examinees exhibiting solution behavior to answer. Alternatively, a unique threshold for each item can be determined through a visual inspection of the distribution of item response times for the specific item (Schnipke, 1995; Wise & DeMars 2006). If the distribution is bimodal, the threshold is set to the first local minimum. Unfortunately, item response times do not always exhibit a bimodal distribution. Another method, referred to as cumulative proportion correct thresholds, integrates both response time and response accuracy to determine the threshold, which is the maximum response time where the proportion of examinees answering the item in that duration or faster scored at chance level (Guo et al., 2016). However, this method is often plagued by complexity. A more practical approach involves setting an item-level Normative Threshold (NT), such as 10% or 20% of the mean response time of all examinees (Wise & Ma, 2012). These diverse methods allow for the identification of rapid guessing behavior on each item. By analyzing an examinee's behavior across items, we can then determine whether they are disengaged or not.

To comprehensively assess test-taking engagement, Wise and Kong (2005) proposed a metric called Response Time Effort (RTE), computed as the proportion of items that a test-taker did not rapidly guess. Examinees with an RTE below 0.90 are commonly flagged as disengaged (Swerdzewski et al., 2011; Wise & Kong, 2005). In a meta-analysis, Silm et al. (2020) found a strong correlation (0.72) between RTE and performance, surpassing the correlation (0.33) between self-reported effort and performance. This suggests that RTE may be a more reliable estimate of engagement than self-report.

Test-Taking Engagement Filtering

Indicators of disengagement can be used to adjust aggregate scores and enhance the validity of low-stakes assessment results through the application of filtering techniques either at the examinee or response level (Alahmadi & DeMars, 2024). Examinee-level filtering involves listwise deletion of response sets from disengaged examinees, such as those with an RTE below 0.90. In contrast, response-level filtering employs a pairwise deletion method, removing only rapid guesses and treating them as missing responses without excluding entire response sets. Consider an examinee who rapidly guesses on 11% of the test items (RTE of 0.89). Examinee-level filtering would exclude all their responses, while response-level filtering would only remove the 11% of responses that were rapid guesses. Although response-level filtering has its advantages, Alahmadi and DeMars (2024) argue that examinee filtering is easier to explain to stakeholders and tends to be more politically acceptable and interpretable.

Predictors of Test-Taking Disengagement

Test-taking engagement may be influenced by the following: personal characteristics, item characteristics, and test-taking characteristics. Personal characteristics relate to the individual taking the test. Notably, gender significantly impacts engagement, with more males likely to rapid guess than females (Rios & Soland, 2022). Additionally, language background (Kroehne et al., 2020) and socioeconomic status (Rios & Soland, 2022) also play roles in test-taking engagement. Item characteristics refer to specific attributes of test items, including item length (Wise, 2006), item position (Rios & Soland, 2022; Wise, 2006), and response type, such as multiple choice or constructed response (DeMars, 2000; Michaelides & Ivanova, 2022). Lastly, examinees' test-taking engagement varies under different test-taking conditions or situational factors. For example, rapid guessing tends to be higher on paper-based tests compared to computer-based tests (Chua & Don, 2013; Kroehne et al., 2020), but the absence of a proctor for computer-based tests can also relate to higher instances of rapid guessing (Kroehne et al., 2020; Wise et al., 2019). Recently, researchers have investigated how the time of day at which a student takes a test influences test-taking engagement (Wise et al., 2010; Wise et al., 2024).

Time of Day

In today's educational landscape, many low-stakes tests are conducted remotely, allowing students greater flexibility in choosing the time of day (TOD) to complete their assessments. This shift has prompted practitioners to consider whether TOD influences test-taking engagement. To date, we identified two studies that investigated the relationship between TOD and test-taking engagement: Wise et al. (2010) and Wise et al. (2024).

In 2010, Wise et al. examined the reading and mathematics sections of a large-scale computer adaptive test administered to students in grades 3 through 9. Students were assigned test start times between 7:00 a.m. and 2:00 p.m. by their local educators, with most students tested two to three times throughout the year. Wise et al. (2010) found no significant relationship between rapid guessing behavior and time of year or day of the week. However, they observed that rapid guessing behavior was significantly more prevalent later in the testing day. The percentage of examinees flagged for disengaged test-taking consistently increased across grade levels and tests. For 9th graders taking the math test, 1.7% were flagged at 7:00 a.m., while 4.6% were flagged at 2:00 p.m. Similarly, for 9th graders taking the reading test, 4.1% were flagged at 7:00 a.m., while 11.6% were flagged at 2:00 p.m. These findings suggest that TOD may predict test-taking engagement. Notably, performance scores on both math and reading tests decreased throughout the day, likely due in part to reduced engagement. In 2024, Wise et al. conducted a similar study using the same large-scale tests for students in grades 2 through 8. Consistent with the original study, they found a significant relationship between TOD and rapid guessing, with rapid guessing being more prevalent later in the day for both the math and reading tests.

TOD Effects in Higher Education Assessment

The impact of TOD on rapid guessing behavior has primarily been studied in K-12 testing contexts. Although no direct research has investigated this relationship in higher education settings, studies on cognitive capabilities and performance can provide insights. Notably, college students exhibit better processing speed (Allen et al., 2008), executive control performance (Allen et al., 2008), and semantic processing (Smith, 1987) in the afternoon. Academic performance also varies throughout the day, with some research suggesting improved performance later in the day and inferior performance in morning classes, with lowest performance in the earliest classes (e.g., 8:00 a.m. start time; Dills & Hernandez-Julian, 2008). However, conflicting findings make it challenging to pinpoint the optimal TOD for performance (Kaur et al., 2021; Onyper et al., 2012).

The shift toward technology-based assessment in higher education has introduced greater flexibility in testing outside the classroom, further distinguishing low-stakes assessments in higher education from those in K-12 settings. Unlike K-12 education, where standardized school hours apply to all students, college students follow individualized daily schedules, making it difficult for institutions to establish fixed assessment times. As a result, it is now common for college students to self-select their testing times and complete assessments remotely (e.g., Alahmadi & DeMars, 2022), whereas K-12 students are typically assigned testing times and complete assessments in school settings (e.g., Wise et al., 2024). While this added flexibility benefits college students, it may come at the cost of testing during suboptimal times for engagement and performance. For example, although many college students have more free time in the evening, they may experience reduced mental stamina after attending classes or working throughout the day. Choosing to test in the evening may be more convenient but could lead to decreased performance compared to morning or afternoon testing. Additionally, the standardized schedules in K-12 education create a highly structured daily routine, with each class or event following the previous one. In contrast, college students often have more dispersed schedules, which may contribute to different patterns of engagement decline as the day progresses.

The Current Study

In the current study, we investigate the relationship between TOD and both engagement and performance among test-takers during a low-stakes higher education assessment. Specifically, we address two primary research questions:

1. To what extent does time of day relate to test-taking engagement?
   Given that the limited literature on TOD and motivation has focused on K-12, we refrain from specifying a hypothesis for potential differences in engagement across different times of day.

2. To what extent does time of day relate to performance?
   Research on the relationship between time of day and performance has yielded conflicting results, especially concerning university students' performance in the morning as compared to the afternoon. Although few studies directly compare performance across morning, afternoon, evening, and nighttime, it is likely that fatigue and reduced mental stamina later in the day impacts performance. Therefore, we hypothesize that student performance will vary between those who test in the morning or afternoon versus those who test in the evening or nighttime.

Method

Participants

We analyzed first-year students from a large mid-Atlantic university who participated in a remote computer-based ethical reasoning assessment (ERA). All first-year students (~4,900) attending the university were mandated to complete a series of assessments prior to arriving at the institution. The battery of assessments was randomly assigned to each student. After employing our inclusion criteria[1], we retained a sample of 1,397 students who completed the ERA. While we did not collect demographic information for our specific sample, institutional data for the undergraduate population at the university indicated 57% of students identified as female and 43% identified as male. Additionally, 74.5% of students identified as White, 7.9% as Hispanic/Latino, 5.1% as Asian, 4.8% as identifying with two or more ethnicities, 4.5% as African American, and 3.2% as another ethnicity.

Achievement Test

The computer-based assessment used in this study consisted of 50 multiple choice items that measured students' understanding of eight key components of ethical reasoning: fairness, outcomes, responsibilities, character, liberty, empathy, authority, and rights (Sanchez et al., 2017). Each item presented an ethical issue or dilemma, and students were required to select the component that best aligned with the prompt. All items had eight options in the same order, corresponding to each of the ethical reasoning components. Only one of the eight options was correct for each item, and items were scored as either correct or incorrect. Students received the ERA as their first assessment in the battery, with no completion time restrictions. Importantly, the ERA was low stakes, as students were explicitly informed that their individual performance did not impact their grade point average (GPA) or transcript.

Rapid Guesses

Item response times were recorded for each student and calculated as the time between a student starting an item and moving on to the following item. Initially, response time thresholds were determined individually by each of the four researchers analyzing the response time distribution, the cumulative proportion correct, and the mean NT10, NT15, and NT20 for each individual item. The estimated interrater reliability between researchers' separate thresholds was high (coefficient alpha= .945). The researchers' collective set of thresholds was used to determine the final rapid guess threshold for each item.

Time of Day

Originally, students were given a 12-day window to complete their assessment, but this was later extended to approximately 1 month. Students had the flexibility to complete the ERA at any TOD during the testing period. Each student's local TOD was recorded through the online testing software (Qualtrics) at the start of the ERA. Subsequently, students were categorized into four general periods: morning (8:00 a.m. - 11:59 a.m.), afternoon (12:00 p.m. - 4:59 p.m.), evening (5:00 p.m. - 9:59 p.m.), and nighttime (10:00 p.m. - 7:59 a.m.). These TOD periods were established based on previous psychological TOD research classifications (e.g., Smith, 1987; Allen et al., 2008) and the structure of a typical workday. Although nighttime had the longest interval, we anticipated that fewer students would be testing during those hours.

Data Analyses

To answer Research Question 1, we conducted a Kruskal-Wallis H test to investigate whether the number of rapid guesses is significantly different across TOD. The Kruskal-Wallis H test was chosen because rapid guessing distributions tend to be positively skewed; many students

[1] We only considered students who consented to research and completed the ethical reasoning assessment in full (without skipping questions) in a typical testing environment.

typically do not engage in rapid guessing at all (Wise et al., 2024). To further investigate Research Question 1 at the test level, we performed a chi-square test of independence to explore whether there was a relationship between TOD and the proportion of students filtered using the RTE cutoff .90.

To address Research Question 2, we conducted a one-way ANOVA to compare performance scores across different times of day. We specifically tested the hypothesis that students who began their test in the morning or afternoon would score higher than those who started in the evening or nighttime. Additionally, to examine the impact of TOD on performance after accounting for motivation, we conducted the same analyses after filtering students at an RTE cutoff of 0.90.

Results

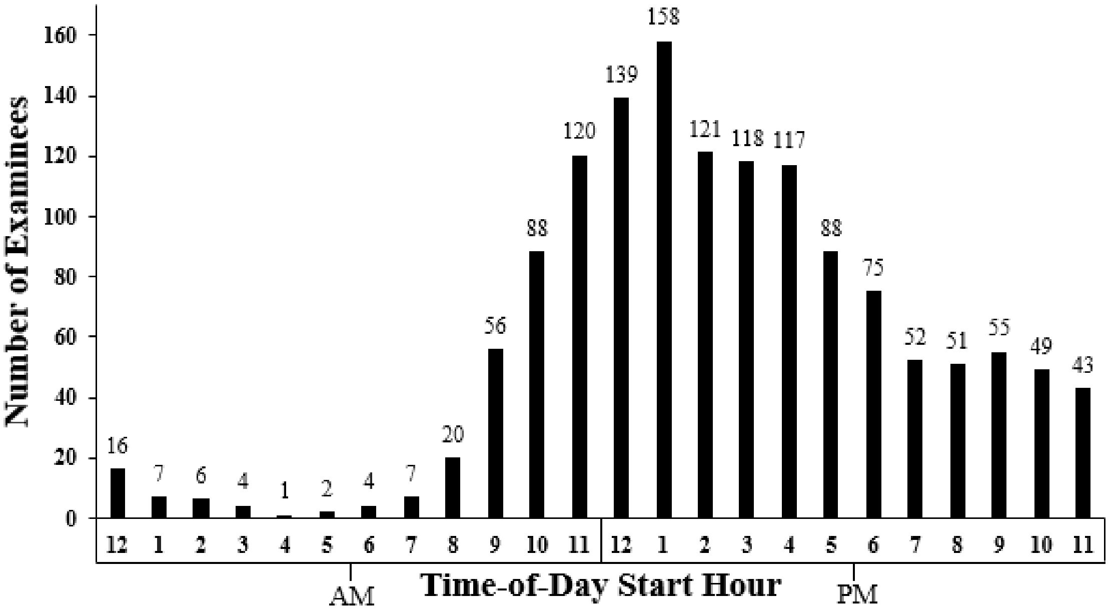
Time of Day of Self-Selected Testing

Students were most likely to self-select to test in the afternoon, with approximately 47% of students starting the ERA between 12:00 p.m. and 4:59 p.m. (see Figure 1). The proportions of students who opted for morning and evening testing were similar, accounting for 20% and 23% of students, respectively. The nighttime group had fewer participants, representing only 10% of students, with 6.5% starting between 10:00 p.m. and 11:59 p.m.

Engagement and Time of Day

Students who began the ERA in the morning or afternoon had one fewer rapid guess, on average, than students who began in the evening or nighttime (see Table 1). Although this finding may have important practical implications, a Kruskal-Wallis H test revealed no statistically significant differences in the number of rapid guesses across the TOD periods ($\chi 2(3)$ = 2.00, p = .571, ηH^2 = 0).

Figure 1. *Frequencies of student time-of-day start hour on the ERA*



Note: The start times for students were recorded in local time zones.

Time-of-Day Effects on Filtering Students

We examined potential differences in student filtering across different times of day using an RTE of 0.90. Students who began the ERA in the evening and nighttime exhibited a higher prevalence of disengagement, resulting in more frequent filtering compared to those who started in the morning and afternoon (refer to Table 2). Approximately 14% of students who tested in the evening and nighttime were filtered out, while 11.6% of students who tested in the morning and afternoon were filtered. A chi-square test of independence conducted between TOD and students filtered at the RTE 0.90 level revealed no significant relationship, $\chi^2(3) = 1.80$, $p = .61$, Cramér's $V = .036$. Although no statistically significant findings emerged, there appears to be a practical association between motivation and TOD, with students exhibiting greater disengagement during the evening and nighttime.

Table 1. *Descriptive statistics for rapid guesses, performance, and filtered performance by time of day.*

Time of Day	N	# of Rapid Guesses		Performance Before Filtering (Out of 50)		Performance After Filtering (Out of 50)	
		M	*SD*	*M*	*SD*	*M*	*SD*
Morning	284	3.00	9.25	30.05	9.73	32.29	7.37
Afternoon	653	3.22	9.33	30.28	9.55	32.53	7.16
Evening	321	3.75	9.95	29.02	9.33	31.52	6.90
Nighttime	139	4.03	10.91	28.65	9.84	30.83	7.59

Table 2. *Percentage test events classified as disengaged for different filtering criteria across day.*

Filtering criteria based on number of rapid guesses	RTE filter threshold	Morning Percent Filtered	Afternoon Percent Filtered	Evening Percent Filtered	Nighttime Percent Filtered
5 or more	.90	11.6	11.6	14.3	13.7

Performance and Time of Day

Before filtering, students who began the ERA in the morning or afternoon, on average, answered one more question correctly than those who started in the evening or nighttime (see Table 1). The mean score of students in the morning and afternoon was significantly higher than the mean score of students in the evening and nighttime, $t(1393) = 2.242$, $p = .025$, [95% CI (0.33, 4.98)], $d = .28$. This finding suggests that student performance is higher during workday hours. After filtering using an RTE cutoff of 0.90, students who took the ERA in the morning or afternoon still scored over one point higher on average than those who started in the evening or nighttime. The difference was significant for both the omnibus one-way ANOVA, $F(3, 1219) = 2.617$, $p = .0496$, $\eta^2 = .006$ and a priori test of the average of morning and afternoon compared to average of evening and nighttime $t(1219) = 2.593$, $p = .010$, [95% CI (0.60, 4.36)], $d = .34$.

Discussion

In this study, we investigated the relationship between test-taking engagement and the TOD when university students chose to complete a low-stakes assessment. Although we did not propose specific hypotheses regarding engagement due to limited prior research, we hypothesized that students testing in the morning and afternoon would perform better than those testing in the evening or nighttime. Our findings revealed a subtle relationship between TOD and engagement: students exhibited greater engagement during morning and afternoon hours compared to evening and nighttime. Additionally, significant performance differences aligned with our hypothesis, with higher scores observed in the morning and afternoon, both before and after adjusting for engagement.

Relationship between Time of Day and Disengagement

The relationship between TOD and student disengagement appears to differ between higher education and K-12 contexts. Among college students, our findings indicated that test-taking engagement remained consistent across the morning and afternoon, whereas previous research with K-12 populations has shown engagement tends to decline in the afternoon (Wise et al., 2010; Wise et al., 2024). This discrepancy may be attributed to the greater scheduling flexibility in college, where students' classes are more spread out, potentially preserving cognitive capacity into the afternoon. Our results also revealed increased disengagement during evening and nighttime hours. This decline may be attributable to fatigue and listlessness becoming more pronounced later in the day, possibly due to the cumulative depletion of energy and cognitive resources.

These findings reflect common student experiences: initiating a test in the evening or nighttime often coincides with reduced energy and a stronger desire to complete tasks quickly, especially as students approach their habitual sleep or social time. In such cases, students may prioritize task completion over sustained, effortful engagement. This behavior aligns with EVT, which suggests that motivation to engage in the test is influenced by perceived value and cost. For low-stakes assessments, the cost of engagement, such as sacrificing sleep or delayed social activities, may outweigh the perceived value, leading to disengagement. Conversely, morning or afternoon testing offers several advantages: students have been awake for fewer hours, retain more cognitive capacity, and face fewer direct costs such as lost sleep or missed social time. These conditions may foster higher levels of engagement, suggesting that scheduling assessments during these periods could support improved student performance and effort.

Relationship between Time of Day and Performance

The lack of performance differences between students who tested in the morning and afternoon aligns with prior research showing mixed findings across these periods (Dills & Hernandez-Julian, 2008; Kaur et al., 2021; Onyper et al., 2012). However, our study revealed a clear decline in performance during the evening and nighttime compared to the morning and afternoon hours. This decline can be partially attributed to increased disengagement, as students were more likely to rapidly guess and answer fewer questions correctly during later testing periods. Notably, even after adjusting for test-taking effort, students who tested in the morning and afternoon consistently outperformed those who tested in the evening and nighttime. Although the mean number of correct responses increased slightly across all four time periods after filtering for effort, the approximate one-question gap between the morning/afternoon and evening/nighttime persisted. This suggests that the self-selected TOD may serve as a proxy for other underlying factors that influence performance beyond what is captured by our engagement measures alone.

Implications, Limitations, and Future Directions

In this study, we aimed to better understand the impact of self-selected TOD on test-taking engagement and performance among college students. While previous research has explored these effects in K-12 settings, differences in administration structure and scheduling flexibility between K-12 and higher education warranted further investigation. As higher education increasingly embraces remote testing, the flexibility it offers has become especially important for non-traditional students, such as adult learners, those returning after gap years, and part-time students, who often face scheduling constraints due to employment or family responsibilities. For these students, the ability to self-select testing times is highly desirable.

However, our findings indicate that testing in the evening is associated with lower engagement and reduced performance. While flexibility remains essential, alternative approaches that preserve choice while optimizing outcomes merits further investigation. For example, universities might consider implementing testing windows that mimic typical workday hours (8 a.m. to 5 p.m.). Yet, restricting testing to these hours could disadvantage students with daytime obligations (e.g., job or family responsibilities). A more balanced solution may involve slightly extending testing hours (e.g., 7 a.m. to 7 p.m.), which could maintain flexibility while mitigating the engagement and performance declines observed during late evening hours. Multi-day or weekend testing options could also support students with weekday conflicts. Although we did not investigate the impact of the day of the week on performance and engagement, prior research suggests this factor is a weak predictor in K-12 populations (Wise et al., 2010). Given the differences in college student demographics and routines, further research is needed before drawing practical conclusions about weekday effects in higher education.

A more practical and student-centered approach may be for universities to recommend, but not require, testing during optimal windows (e.g., 8 a.m. to 5 p.m.). Framing this guidance as evidence-based could encourage students to self-select times that support better performance and engagement. Communicating the rationale behind these recommendations may also enhance perceived value of the assessment, signaling institutional care for student success and outcomes.

These results have important implications for longitudinal studies of curricular effectiveness. If students choose different testing times during pre- and post-assessments, observed performance changes may reflect TOD effects rather than learning gains. For example, students who completed the ERA in our study will be retested later in their academic careers. It will be valuable to examine whether their self-selected testing times remain consistent or vary as such differences could influence the validity of longitudinal score interpretations.

Despite observing practical differences in performance across different TODs, these disparities persisted even after accounting for engagement. The self-selection of testing times limits experimental control and introduces potential confounding variables. Beyond engagement, factors such as fatigue and listlessness likely contribute to performance declines, as previously demonstrated (Weirich et al., 2017). These effects may be moderated by individual characteristics, including gender, employment status, family responsibilities, and socioeconomic status. For example, a student working full-time from 9:00 a.m. to 5:00 p.m. may only be able to test in the evening, potentially leading to lower engagement and performance compared to a peer with more flexible availability. Future research should explore these individual differences to better understand the mechanisms behind TOD-related performance declines and inform equitable testing practices.

Overall, our findings highlight the complex relationship between TOD and assessment outcomes, suggesting that factors such as fatigue, daily schedules, and personal responsibilities significantly influence engagement and performance. Recognizing and accommodating these influences is essential for optimizing testing conditions and improving the validity of academic assessments.

Conclusion

Our study demonstrates that test-taking engagement varies across TOD, and that TOD can significantly influence performance. Students who choose to take tests in the evening or nighttime tend to disengage and achieve lower scores. These findings suggest that institutions should consider limiting evening and nighttime testing to enhance student performance and engagement. However, because this study did not employ experimental controls, TOD may act as a proxy for other underlying characteristics that influence outcomes.

Our research contributes valuable insights into how TOD relates to effort and performance among university students—a population that has been underexplored in this context. By highlighting performance variations based on TOD, our study lays the foundation for future investigations into how flexible testing schedules may affect academic outcomes assessment results. These findings also prompt important considerations for educational policy, particularly regarding the scheduling of low-stakes assessments and implications of allowing students to self-select their testing times.

Overall, this study advances our understanding of TOD's relationship with test-taking engagement and performance in higher education. It also paves the way for further research into the individual and contextual factors that shape these outcomes, with potential implications for assessment design, scheduling practices, and the interpretation of scores in flexible testing environments.

References

Alahmadi, S., & DeMars, C. E. (2022). Large-scale assessment during a pandemic: Results from James Madison University's remote Assessment Day. *Research & Practice in Assessment*, *17*(1), 5–15.

Alahmadi, S., & DeMars, C. E. (2024). Comparing examinee-based and response-based motivation filtering methods in remote low-stakes testing. *Applied Measurement in Education*, *37*(1), 43–56. https://doi.org/10.1080/08957347.2024.2311927

Allen, P. A., Grabbe, J., McCarthy, A., Bush, H., & Wallace, B. (2008). The early bird does not get the worm: Time-of-day effects on college students' basic cognitive processing. *The American Journal of Psychology, 121*(4), 551–564. https://doi.org/10.2307/20445486

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (Eds.). (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Chua, Y. P., & Don, Z. M. (2013). Effects of computer-based educational achievement test on test performance and test takers' motivation. *Computers in Human Behavior 29*(5), 1889–1895. https://doi.org/10.1016/j.chb.2013.03.008

DeMars, C. E. (2000). Test stakes and item format interactions. *Applied Measurement in Education*, *13*(1), 55–77. https://doi.org/10.1207/s15324818ame1301_3

Dills, A. K., & Hernandez-Julian, R. (2008). Course scheduling and academic performance. *Economics of Education Review 27*(6), 646–654. https://doi.org/10.1016/j.econedurev.2007.08.001

Eccles, J. S., Adler, T. F., Futterman, R., Goff, S. B., Kaczala, C. M., Meece, J. L., & Midgley, C. (1983). Expectancies, values, and academic behaviors. In J. T. Spence (Ed.), *Achievement and achievement motivation* (pp. 74–146). San Francisco, CA: W. H. Freeman.

Eccles, J. S., & Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annual Review of Psychology*, *53*(1), 109–132. https://doi.org/10.1146/annurev.psych.53.100901.135153

Gorgun, G., & Bulut, O. (2023). Incorporating test-taking engagement into the item selection algorithm in low-stakes computerized adaptive tests. *Large-scale Assessments in Education, 11*(27). https://doi.org/10.1186/s40536-023-00177-5

Guo, H., Rios, J. A., Haberman, S., Liu, O. L., Wang, J., & Paek, I. (2016). A new procedure for detection of students' rapid guessing responses using response time. *Applied Measurement in Education, 29*(3), 173–183. https://doi.org/10.1080/08957347.2016.1171766

Kaur, P., Kumar, H., & Kaushal, S. (2021). Affective state and learning environment based analysis of students' performance in online assessment. *International Journal of Cognitive Computing in Engineering*, *2*, 12–20. doi:10.1016/j.ijcce.2020.12.003

Kroehne, U., Deribo, T., & Goldhammer, F. (2020) Rapid guessing rates across administration mode and test setting. *Psychological Test and Assessment Modeling 62*(2), 147–177. https://doi.org/10.25656/01:23630

Michaelides, M. & Ivanova, M. (2022). Response time as an indicator of test-taking effort in PISA: country and item-type differences. *Psychological Test and Assessment Modeling*, *64*(3), 304–338.

Onyper, S. V., Thacher, P. V., Gilbert, J. W., & Gradess, S. G. (2012). Class start times, sleep, and academic performance in college: A path analysis. *Chronobiology International, 29*(3), 318–335. https://doi.org/10.3109/07420528.2012.655868

Pastor, D. A., Ong, T. Q., & Strickman, S. N. (2019). Patterns of solution behavior across items in low-stakes assessments. *Educational Assessment, 24*(3), 189–212. https://doi.org/10.1080/10627197.2019.1615373

Rios, J. (2021). Improving test-taking effort in low-stakes group-based educational testing: A meta-analysis of interventions. *Applied Measurement in Education, 34*(2), 85–106. https://doi.org/10.1080/08957347.2021.1890741

Rios, J. A. & Deng, J. (2021). Does the choice of response time threshold procedure substantially affect inferences concerning the identification and exclusion of rapid guessing responses? A meta-analysis. *Large-Scale Assessments in Education*, *9*(1), 18. https://doi.org/10.1186/s40536-021-00110-8

Rios, J. A., Guo, H., Mao, L., & Liu, O. L. (2016). Evaluating the impact of careless responding on aggregated-scores: To filter unmotivated examinees or not? *International Journal of Testing 17*(1), 74–104. https://doi.org/10.1080/15305058.2016.1231193

Rios, J. A., & Soland, J. (2022). An investigation of item, examinee, and country correlates of rapid guessing in PISA. *International Journal of Testing*, *22*(2), 154–184. https://doi.org/10.1080/15305058.2022.2036161

Sanchez, E. R. H., Fulcher, K. H., Smith, K. L., Ames, A., & Hawk, W. J. (2017). Defining, teaching, and assessing ethical reasoning in action. *Change: The Magazine of Higher Learning, 49(*2), 30–36. https://doi.org/10.1080/00091383.2017.1286215

Schnipke, D. L. (1995). *Assessing speededness in computer-based tests using item response times* [Unpublished doctoral dissertation]. Johns Hopkins University.

Silm, G., Pedaste, M., & Täht, K. (2020). The relationship between performance and test-taking effort when measured with self-report or time-based instruments: A meta-analytic review. *Educational Research Review, 31,* 100335. https://doi.org/10.1016/j.edurev.2020.100335

Smith, A. P. (1987). Activation states and semantic processing: A comparison of the effects of noise and time of day. *Acta Psychologica, 64*(3), 271–288. https://doi.org/10.1016/0001-6918(87)90012-6

Swerdzewski, P. J., Harmes, J. C., & Finney, S. J. (2011). Two approaches for identifying low-motivated students in a low-stakes assessment context. *Applied Measurement in Education, 24*(2), 162–188. https://doi.org/10.1080/08957347.2011.555217

Weirich, S., Hecht, M., Penk, C., Roppelt, A., & Böhme, K. (2017). Item position effects are moderated by changes in test-taking effort. *Applied Psychological Measurement, 41*(2), 115–129. https://doi.org/10.1177/0146621616676791

Wise, S. L. (2006). An Investigation of the differential effort received by items on a low-stakes computer-based test. *Applied Measurement in Education*, *19*(2), 95–114. https://doi.org/10.1207/s15324818ame1902_2

Wise, S. L. (2020) Six insights regarding test-taking disengagement. *Educational Research and Evaluation*, *26*(5–6), 328–338. https://doi.org/10.1080/13803611.2021.1963942

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment*, *10*(1), 1–17. https://doi.org/10.1207/s15326977ea1001_1

Wise, S. L., & DeMars, C. E. (2006). An application of item response time: The effort-moderated IRT model. *Journal of Educational Measurement*, *43*(1), 19–38. https://doi.org/10.1111/j.1745-3984.2006.00002.x

Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education*, *18*(2), 163–183. https://doi.org/10.1207/s15324818ame1802_2

Wise, S. L., Kuhfeld, M. R., & Lindner, M. A. (2024). Don't test after lunch: The relationship between disengagement and the time of day that low-stakes testing occurs. *Applied Measurement in Education, 37*(1), 14–28. https://doi.org/10.1080/08957347.2024.2311925

Wise, S. L., Kuhfeld, M. R., & Soland, J. (2019). The effects of effort monitoring with proctor notification on test-taking engagement, test performance, and validity. *Applied Measurement in Education*, *32*(2), 183–192. https://doi.org/10.1080/08957347.2019.1577248

Wise, S. L., & Ma, L. (2012, April). Setting response time thresholds for a CAT item pool: The normative threshold method [Paper presentation]. National Council on Measurement in Education 2012 Annual Meeting, Vancouver, BC, Canada.

Wise, S. L., Ma, L., Kingsbury, G. G., & Hauser, C. (2010, May). An investigation of the relationship between time of testing and test-taking effort [Paper presentation]. National Council on Measurement in Education 2010 Annual Meeting, Denver, CO, United States.