An Analysis of Students’ Makeup-Exam Results and Performance in Undergraduate Biology Courses

DENNIS D. TARASI

ABSTRACT

A consistent challenge for undergraduate instructors is how to properly and objectively assess students who cannot attend regularly scheduled exams. Though many alternatives exist, perhaps the most common strategy is to allow students to take a makeup exam at a different time. Many instructors avoid this option for fear of the students gaining an unfair advantage in their exam preparations. I assessed student performance on makeup exams in relation to their typical performance on other exams throughout the course, to determine if there was any signal of score improvement or decline on makeup exams. I analyzed the data in regard to when students took the makeup exam, what their excuse for missing the regularly scheduled exam was, and the type of course in which they took the makeup. Students’ makeup-exam scores were not significantly different from their regular-exam scores, though students who took a makeup due to a school-sponsored activity scored worse than students taking a makeup due to family emergencies or illness. While this research cannot definitively state that makeup exams do not provide some sort of advantage to student scores, it suggests that if students are trying to "game the system," at the very least, they aren’t winning.

Key Words: makeup exam; undergraduate courses; exam performance; missed exam; assessment; assessment of student learning; biology.

Introduction

A significant and imposing challenge in undergraduate education is to provide equitable instruction and assessment to all students, such that students’ performance is considered as objectively as possible for the field and in relation to their peers. Student behaviors often complicate this challenge, particularly regarding exams. In theory, regularly scheduled exams are practical and objective measures of student mastery of course content. In practice, students regularly miss these assessments and muddy the waters of objectivity regarding their course success. Students miss both classes and regularly scheduled exams, with excuses ranging from entirely valid to wildly improbable (Song, 2013). These excuses often accumulate in courses with lenient makeup-exam policies, as students more regularly miss exams throughout a semester, though they rarely miss the final exam (Abernathy & Padgett, 2010). Regardless of makeup-exam format or course level, students taking makeups generally score lower than their peers on all exams and final course grades, though the direction of causation remains unclear (Kahn, 1995, 2000). We cannot say for certain whether poorly performing students ask for makeups more often in an effort to gain a better chance at success, or if makeups disadvantage students throughout a course and thus result in a lower course grade.

A wide array of makeup policies, with corresponding levels of efficacy and objectivity, exists throughout the education literature. These options include dropping the lowest exam score, double-counting the final exam (Anderson, 1981), providing a comprehensive makeup exam (Kottke, 1985), having an optional final exam that becomes mandatory for those who miss an exam (Buchanan & Rogers, 1990), and allowing students who miss an exam to develop and answer their own exam questions to demonstrate their mastery (Carper, 1995). Each strategy in handling student absences from exams presents its own challenges, including the logistics of implementation, academic integrity of the students, perceived lack of fairness to aggrieved students, and accurate and objective assessment of student mastery. Often, the rationale for substituting scores or dropping the lowest exam score, instead of providing the opportunity for a makeup exam, is to prevent students who take a makeup from being unfairly advantaged compared to their peers who took the exam at the regularly scheduled time. Instructors are concerned that students may “game the system” by delaying their exam time to have more time to study and learn about exam content from their peers, unfairly inflating their scores in relation to their typical performance.

Unfortunately, changing the exam may actually unfairly disadvantage students who miss the exam in relation to the rest of the class (Funk & Dickson, 2011). Dropping the lowest exam score or substituting a comprehensive final for the missed exam does not expose students to assessment items related to that content in the same fashion as their peers, which can skew seemingly objective assessments and harm final exam scores due to that reduced exposure. With the ultimate goal of objectively measuring student mastery, it seems that changing the
format of the exam does more harm than good. As a result, many instructors have opted to provide makeup exams that are of identical or very similar format and content as the regular exam, offered shortly before or after the regular exam. While this option inherently presents increased logistical hurdles for instructors in comparison to dropping or substituting exams, instructors who provide full makeup consider this the fairest and most objective way to assess students and expose them to key course content. The limited scholarly research on the efficacy of makeup exams causes instructors and departments to rely on their own anecdotal perceptions and preferences concerning best practices regarding makeup exams. Kahn (1995) determined that introductory psychology students scored worse on makeup exams than on regularly scheduled exams, which was corroborated among upper-level psychology students (Kahn, 2000). It is not clear, however, how generalizable these results may be for other fields, levels of study, or types of institution.

I have personally maintained a very flexible makeup-exam policy, sometimes asking for a doctor’s note or Googling a loved one’s obituary to verify an excuse, but otherwise allowing students to take the makeup exam when their schedule allowed within a reasonable time frame around the exam. My anecdotal experience suggests that students do not generally perform better on makeup exams than normal, so the logistical challenge of setting up a time during office hours for them to complete the exam does not outweigh any potential score benefit they may incur by taking the makeup. With no true research support in my discipline about makeup-exam performance versus regular exams, I chose to investigate.

Here, I address several components of a common makeup-exam strategy in undergraduate biology instruction: allowing students to take a similarly or identically formatted exam shortly before or after the regularly scheduled exam time. I consider three distinct questions about makeup exams:

1. How does taking the exam at a different time from the regularly scheduled time impact a student score on that exam, compared to their performance on other exams?
2. Does student performance on makeup exams vary with the reason provided for missing the regularly scheduled exam (illness, family emergency, school activity, etc.)?
3. Does student performance on makeup exams vary with the type of course (general education, service to other majors, biology majors) or typical achievement level of the student (regular-exam performance, compared to their peers)?

I generally expected the answers to these questions to match the null hypothesis that makeup-exam performance would not be significantly different from regular-exam performance for any type of student, excuse, or course. That is, I expected that students who took a makeup exam would do as well (or as poorly), in comparison to their peers, on the makeup exam as on their regularly scheduled exams, regardless of the type of course, the reason for taking a makeup, or the timing of the makeup.

Methods

Data Collection

I collected exam data for 73 makeup exams from 22 sections of eight courses offered in the Department of Biology at St. Ambrose University (SAU) in the 2018–19 academic year. Located in Davenport, Iowa, SAU is a private, comprehensive university grounded in the liberal arts. With over 60 undergraduate majors and 15 graduate programs, SAU enrolls ~2300 undergraduates and ~750 graduate students. Courses represented the breadth of the typical Biology Department undergraduate offerings, including general education courses meeting the liberal arts requirements, service courses for other majors at the university (primarily nursing, exercise science, and human performance and fitness), and courses required to complete the undergraduate biology major. In these courses, instructors offered individual, handwritten lecture exams with the opportunity for makeup exams if a student missed the regularly scheduled exam time. The makeup exams could be taken before or after the regular offering and were of similar length, format, and content as the regular exam. Many of the makeup exams were identical to the regular exam, and all were of comparable rigor. This course policy is not universal throughout the department, so some instructors’ policies (dropping lowest exam score, replacing with the final, etc.) excluded their courses from this analysis. Instructors noted makeup exams in their gradebooks with both an estimate of the number of hours before or after the regularly scheduled exam that the makeup exam was taken and the reason why a student missed the regularly scheduled exam. Reasons for missing the exam were assigned to one of four categories:

1. School-sponsored activity (athletic event, travel for band/theater, military training, etc.)
2. Family emergency or funeral
3. Student illness (including mental health episodes)
4. Other (graduate school visits, jury duty, medical appointments, etc.)

If students did not provide an excuse for their absence, or if instructors did not record an excuse in the data, these data were assigned to the “other” category. These four categories were chosen because they generally represented different levels of wellness and states of mind for the students as they took the exams.

At the end of the semester, each instructor provided me with an anonymous spreadsheet of all exam scores from that section for students who completed the course (IRB exempt, SAU case no. 2020329). Students who dropped the course or walked away before the final were not included, and optional finals were not included in the data set. The compiled data represent 73 makeup exams from 60 students out of a total 413 completed enrollees across all courses. Courses were categorized as primarily one of three categories: general education courses meeting the liberal arts requirements, service courses for other majors at the university (primarily nursing, exercise science, and human performance and fitness), and courses required to complete the undergraduate biology major (Table 1).

Multiple students in several service courses completed multiple makeup exams in a given semester. Note that because many students are enrolled in multiple biology courses in a given semester or year, individual students may count as an enrollee in more than one distinct course section. Because some instructors in the department follow different makeup-exam policies, or because some courses had no students complete a makeup exam, a significant portion of the department’s exam scores were not included for analysis.
Table 1. Course enrollments and number of makeup exams within the data set. The clear majority of course enrollees within the data set, and taking makeup exams, were from service courses.

<table>
<thead>
<tr>
<th>Primary Demand for Course</th>
<th>Courses Included in This Analysisa</th>
<th>Students Completing All Exams</th>
<th>Students With at Least One Makeup Exam</th>
<th>Total Makeup Exams</th>
</tr>
</thead>
<tbody>
<tr>
<td>General education</td>
<td>• BIOL 109: Environmental Science (4)</td>
<td>93</td>
<td>14</td>
<td>14</td>
</tr>
<tr>
<td>Service to other majors/programs</td>
<td>• BIOL 101: Principles of Biology (4) • BIOL 211: Introductory Microbiology (2) • BIOL 230: Human Anatomy and Physiology I (6) • BIOL 232: Human Anatomy and Physiology II (2)</td>
<td>247</td>
<td>38</td>
<td>51</td>
</tr>
<tr>
<td>Courses for majors</td>
<td>• BIOL 200: General Biology II (1) • BIOL 201: Diversity of Living Systems (1) • BIOL 303: Genetics (1)</td>
<td>73</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>413</td>
<td>60</td>
<td>73</td>
</tr>
</tbody>
</table>

aThe number of sections included is in parentheses.

Data Analysis

For each course section, I calculated the mean and standard deviation of each class set of exams for all students who took every exam at the regularly scheduled time. In courses where any student took makeup exams, all students with any makeup were omitted from the class-wide analysis for all exams, so the role of the makeup exam would not impact the overall course-wide exam scores. Because sections and instructors varied in their course means and standard deviations, I calculated z-scores to standardize a student’s performance, in comparison to that of their peers, across sections and instructors. The z-score can be calculated using the equation

\[ z = \frac{x - \mu}{\sigma} \]

where \( x \) is the student score for a given exam, \( \mu \) is the class mean for the regularly scheduled exam, and \( \sigma \) is the class standard deviation for the regularly scheduled exam. Both regular-exam and makeup-exam score distributions were normal in nature, ensuring that z-scores are a suitable metric of evaluation.

For any student who completed a makeup exam in that section, I calculated the z-score for each of that student’s exams for the semester. A positive z-score indicated that a student performed better on the exam than the class mean, while a negative z-score indicated poorer performance than the class mean. The further a z-score is from zero, the greater the difference between the student score and the class mean.

I compared the z-score of each student’s makeup exam to the mean z-score of their regularly scheduled exams, in an effort to determine if taking the makeup exam changed their exam performance in comparison to that of their peers. For students who took multiple makeup exams, each makeup-exam z-score was compared to the regular-exam z-score. I determined the number of makeup exams that scored higher than the typical exam score, for all exams as well as for several subsets, matching the research questions: whether the makeup exam was completed before or after the regular exam, the rationale provided for the makeup, the target population of the course, and the achievement level of the students taking the makeup exams (high achievers had a positive z-score on regularly scheduled exams, while low achievers had a negative z-score on regularly scheduled exams).

I used a one-way analysis of variance (ANOVA) for all makeup-exam students to determine if makeup-exam z-scores were statistically different from regularly scheduled z-scores (\( \alpha < 0.05 \)). I then partitioned the entire data set for analysis via one-way ANOVA based on subsets related to the research questions: makeup-exam timing, the rationale provided for the makeup, the target population of the course, and student achievement level. I assessed the pairwise statistical significance for all groups within each partition based on their regular-exam z-score, makeup-exam z-score, and change in z-score from regular to makeup exams.

Students taking multiple makeup exams provide a very interesting test case, with multiple data points, about how makeup-exam performance relates to regular-exam performance. A total of 11 students across six course sections completed multiple makeup exams, representing 24 total makeup exams. I assessed these students in comparison to those taking single makeup exams using one-way ANOVA.

Results

Across all 73 exams, 40 makeups demonstrated an improvement in student z-score compared to the regular exams, and 33 demonstrated a decline in z-score (Figure 1). Early and late exam-takers demonstrated similar overall trends in z-score change. High-achieving students (those with positive regular-exam z-scores) had 14 improvements and 15 declines, contrasting with low-achieving students (26 improving, 18 declining). These findings collectively indicate a slight regression toward the mean; that is, students’ makeup scores were closer to average than their regular-exam scores, in comparison to their peers.

Students taking a makeup exam due to a school-sponsored activity demonstrated a marked decline in exam performance; six exams improved compared to the regular z-score, while 13 declined. For all other reasons for a makeup, makeup-exam performance was typically higher than regular-exam performance. Finally, students in
general education courses were as likely to improve as to decline on their makeup exams in comparison to their regular-exam performance, while those in majors’ courses and service courses matched general trends, with more makeup exams improving than declining in comparison to regular-exam performance.

The mean regular-exam $z$-score of makeup students was $-0.320$, indicating that the typical student taking a makeup exam was slightly below average for their course enrollment (Figure 2). The mean makeup-exam $z$-score was $-0.266$, which was not significantly different from the regular-exam scores for these students. These data corroborate both the small score improvement of most students and the regression toward the mean mentioned previously. These data match the trend of median $z$-score values ($-0.26$ for regular exams, $-0.15$ for makeup exams), suggesting a limited skew due to outliers. Students taking the makeup exam early had a significantly higher regular-exam $z$-score ($0.136$) than students taking the makeup exam late ($-0.428$), represented by a $P$-value $<0.05$. There was no significant difference in makeup-exam scores ($P = 0.14$) or in the change in $z$-score for these two groups ($P = 0.61$). The amount of time before or after the regular exam time was not predictive of student scores beyond the categorical measure mentioned above, generally supporting the null hypothesis.

High-achieving students were, unsurprisingly, significantly stronger on both regular exams ($0.494$) and makeup exams ($0.395$) than low-achieving students ($-0.857$ and $-0.701$, respectively). While both groups demonstrated some regression toward the mean, neither change was statistically significant for high achievers ($P = 0.44$), for low achievers ($P = 0.33$), or among groups ($P = 0.1$), generally supporting the null hypothesis.

Students taking a makeup exam due to a school-sponsored activity demonstrated a marked decline in $z$-score ($-0.324$), which was statistically significant in comparison to the increase in score for family emergencies ($0.264$) and illness ($0.287$). None of the regular-exam $z$-scores or makeup-exam $z$-scores were significantly different from one another between or within groups (lowest $P$-value among all measures equals $0.20$). The change in regular-exam and makeup-exam $z$-scores for the “other” category was not significantly different from any of the categories. These findings, particularly the significance of the activity group’s change, rejected the null hypothesis.

There was no significant difference between or within groups related to the target audience of the course (lowest $P$-value among all measures equals $0.15$). The average student taking a makeup in all three course types was relatively below average, though not

Figure 1. Number of makeup exams with performance, in relation to that of peers, that improved (light gray) or declined (dark gray) compared to a student’s performance on regularly scheduled exams ($n = 73$). (A) More students improved than declined across all exams, regardless of when the makeup was taken. (B) High-achieving students experienced similar numbers of improvements and declines in makeup exam performance, while low-achieving students improved more than they declined. (C) Students missing for an activity declined more than they improved, while all other reasons for missing saw more improvements than declines. (D) General education students matched improvements and declines, while majors and service courses saw more improvements than declines.
significantly so. Each course type demonstrated a nonsignificant regression toward the mean from regular exams to the makeup exam. These findings supported the null hypothesis.

Students taking multiple makeup exams earned regular-exam z-scores (−0.491) that were qualitatively lower, but not significantly different from those of students who took only one makeup exam (−0.239). Among the 24 exams taken, students improved on their regular z-score on exactly half of the exams. Four students improved on both of the makeup exams they took, and three declined on both of their makeup exams. Both students who took three makeup exams across the semester improved their score, compared to their regular z-score, on one of the three exams. The mean improvement from the regular z-score to the makeup for the 24 exams was nonsignificant (0.008).

**Discussion**

This research demonstrates that students taking makeups generally do not score as well on all exams as their peers, corroborating previous research. Kahn (1995) determined that introductory psychology students who took makeup exams generally scored lower than their peers on all exams, which was also the case in upper-level psychology courses (Kahn, 2000) and which matches my findings. In both studies, however, Kahn determined that students generally performed better on regular exams than on makeup exams, which contradicts my findings. Students’ performance on makeups was not significantly different from their regular-exam performance in my data set, despite a qualitative improvement, suggesting that students taking a makeup are not earning an unfair advantage over their peers, nor are they facing undue hardship. This confirms my anecdotal experience, and that of many colleagues, supporting the use of the makeup as an objective tool to ensure consistent evaluation of students across a course. This also assists the previous research in determining a direction of causation regarding makeups and student performance. Lower-achieving students tend to request makeups more than the rest of the class, as evidenced by the below-average regular-exam performance of the mean student taking makeups in my data set and the greater
number of regular-exam z-scores below zero. However, these students did not demonstrate a statistically significant change in score, which shows that makeup exams are not unfairly disadvantaging (or advantaging) student performance.

While most makeup exams demonstrated some general improvements in score and an overall regression toward the mean regardless of regular-exam score, these improvements were relatively minor and not statistically significant. Although the lack of significance may be due to the relatively small sample size, the 0.06 z-score improvement across all exams is equivalent to <1% of a normal exam-score improvement, where standard deviation in a normal distribution is 10 percentage points. For example, the data indicate that students who take makeups typically earn a 72% on regularly scheduled exams (class average = 75% with a normal “bell curve” distribution). That same student taking a makeup exam would be likely to earn a 73%. Thus, the typical change in makeup score does not significantly impact a student’s course grade in a way that should deter instructors from using makeups in the future. This aligns with my anecdotal experience that students taking exams later are likely not spending more time studying or gaining unfair content advantages from peers, but are simply delaying when they study until a different day. Thus, their scores do not generally improve.

High achievers and early test-takers were the only groups of makeup students with positive regular-exam z-scores. It stands to reason that students would generally opt for early makeup exams only if they feel confident in their abilities and preparation for the exam. With flexible makeup policies, students could opt to take the exam late to gain more time for preparation. Both high achievers and early makeups demonstrated a general decline in makeup-exam z-score. The lack of significance may be due to small sample size, so it is worth considering if these trends are indicative of underlying psychological changes. Just as low-performing students may benefit very slightly from increased time and preparation, successful students may be slightly harmed by a change in routine associated with makeup exams. Because they are taking a makeup exam, these students may also be more preoccupied with other thoughts and obligations than during their normal test prep.

I was surprised to learn that students taking makeups due to school-sponsored activities experienced the greatest decline in z-score in relation to their regular performance. I had expected those with family emergencies or funerals to be most affected in their preparations and emotional well-being. I was not surprised that the illness reason generally saw improved scores, as students can recover and prepare with little emotional burden or impact. It may stand to reason that important athletic events or performances would weigh heavily on students’ minds and schedules, such that their preparations suffer before and after the event. I wondered if perhaps the activity group declined due to confounding factors of makeup-exam timing, of the three major reasons, students with activities were the only ones to take the makeup exam early (seven of 19 took the exam early, none of which were on the day of the exam). Given that early makeups correspond to a reduced opportunity to reinforce and review content, ask questions, and prepare, perhaps this reduced student performance within this group. However, eight of the 12 activity students taking the exam late also scored lower on the makeup, suggesting that the general impact of the activity on student performance extends beyond timing.

### Limitations

This research represents a wide array of makeup exams across a relatively large department within our university. Unfortunately, because makeup exams are not entirely common, these statistical analyses are limited by small sample sizes, particularly in partitioning the data according to reason for missing and type of course. While the sampled data set is representative of the overall enrollment of students within our department offerings (primarily service courses, with a balanced representation of upper- and lower-level offerings), many of these groups were represented by only a handful of makeup exams. In particular, several upper-level courses within the biology major are project-based, so there are fewer exams to compare for that population. There were still several upper-level courses included in this data set (see Table 1). There would likely be more statistically significant differences with more data points, but it is unclear how much general trends would change.

This research represents an exploratory study within one department of one type of higher-learning institution. Because of the nature of the given exams, and the variety of target audiences within these courses, it is possible that these conclusions extend beyond this sample group to other types of undergraduate departments or institutions. However, without further support from additional research, inferences are not intended to generalize beyond this sample.

It is important to recognize that the makeup exams given still did not fully match the conditions of the regular exam. Students took the exam in a different setting, at a different time of day (for better or worse), and with a different level of preparation and state of mind than if forced to take the regularly scheduled exam. Perhaps students would have scored even worse without a makeup option, so the makeup improved their grade in relation to what they would have earned. We cannot rule out this possibility. At the very least, students are not dramatically improving their grades by taking a makeup, compared to their normal performance.

To be included in this data set, instructors did not have to verify the validity of students’ excuses (requiring doctor’s notes, seeing obituaries, etc.), though many instructors did so. Thus, students may be incentivized in these courses to provide dishonest excuses, thereby allowing themselves more time to prepare for the exam. Such invalidated excuses may skew the results of this study, particularly when considering the classification of different types of excuses. The “other” group could refer to a wide array of states of mind, including a misclassified category from another excuse. To my knowledge, only one makeup exam was listed without an excuse provided from the instructor, so the “other” category should fulfill its defined purpose without significant concern of skewing. Fortunately, school-sponsored activities were the only excuses to provide a significant difference from the other provided excuses. I am inclined to believe that those excuses were most likely valid, such that the overall conclusions of the paper should remain sound.

This study does not consider the magnitude of change in makeup-exam performance in comparison to regular-exam performance, but merely whether students improved or declined in comparison to their peers. While I did evaluate mathematical changes with means and one-way ANOVAs, further examination of larger data sets could elucidate which types of student scores change, and by how much. I also did not distinguish exams given at different times of the semester. Perhaps students who miss early exams fall
behind the rest of the class in foundational skills as they prepare for the makeup exam, thus disadvantaging their regular-exam scores for the rest of the semester. Perhaps students who miss late exams are not fully prepared for the most complex and difficult of course topics, thus lowering their makeup-exam performance. These questions could make for enlightening future studies to expand beyond this preliminary analysis.

○ Conclusion

Overall, this research demonstrates that students taking makeup exams are not gaining a significant advantage over their peers. Thus, if the logistical challenges of scheduling the makeup can be overcome, I believe that the benefits of providing students the exposure to unit-specific assessment items comparable to that given their peers, as well as another objective data point of their performance, outweigh any concerns of overarching impacts on an individual’s grade. This system of makeup exams also levels the playing field with a student’s peers, such that if a given exam is easier or more difficult than the others, or if an exam is in an area that aligns closely with the student’s strengths, the student has the opportunity to perform appropriately regardless of the exam’s rigor, compared to the other exams given throughout the course. If student exams are dropped or a different makeup is given, these methods may skew the resulting course grades significantly based on the relative rigor of a given exam compared to others.

A common concern voiced among faculty who are hesitant to provide students with makeup exams is that students may simply be trying to “game the system,” such that taking the exam at a later date allows them to earn a higher score than they would have if they’d taken the exam at the regularly scheduled time. But as Kahn (1995, 2000) noted, students’ makeup-exam scores are not significantly better than their regular-exam scores. My findings corroborate Kahn’s conclusions in a different undergraduate discipline. Coupled with those previous findings, this research demonstrates that if students truly are trying to game the system by using makeup exams, at the very least, they aren’t clearly winning.

○ Suggestions for Future Study

This research demonstrates an initial exploration of just one type of makeup exam within one type of undergraduate department. Future research can extend these conclusions by evaluating student performance on other types of makeup exams (e.g., double-counting performance on the relevant sections of a final exam, student-designed makeup exams). It would also be relevant to consider how makeup exams in biology compare to other exam-intensive disciplines (e.g., nursing, engineering, chemistry) or at different levels of study (e.g., introductory biology majors vs. upper-level courses). Other studies could evaluate the rigor of the exams, to determine how students perform on questions at varying levels of Bloom’s taxonomy, or the difficulty levels of particular units within a course. Finally, it would be interesting to look at whether makeup exams benefit certain groups or types of students more than others.

○ Acknowledgments

I am very grateful to many colleagues in the SAU Department of Biology who assisted with data collection throughout the 2018–19 academic year and provided valuable comments during manuscript development. These colleagues included Drs. Neil Aschliman, Amy Blair, Matthew Halfhill, Kirk Kelley, Shannon Mackey, Brenda Peters, and Anita Zahs. I thank the anonymous reviewers who provided valuable comments during review of the manuscript.

References


DENNIS D. TARASI is an Assistant Professor of Biology at St. Ambrose University, Davenport, IA 52803; e-mail: tarasidennisd@sau.edu.