Assessing Long-Term Effects of COVID-19–Induced Grading Policy Changes on Voluntary S/U Behavior

Sophia Yang

Acknowledgements

I am deeply grateful to Professor Mine Çetinkaya-Rundel, my advisor, instructor, and Director of Undergraduate Studies in Statistical Science whose guidance, mentorship, and support shaped every stage of this project. I would also like to thank Professor Edwin Iversen and Professor Jerome Reiter for serving on my defense committee, instructing me during my undergraduate studies, and providing the statistical expertise and insights that informed many of the methods used in this thesis. I am especially grateful to Dr. Jennifer Hill for providing the dataset that made this research possible, as well as for your subject-matter expertise and guidance on the framing of research questions and ethical considerations. Finally, I extend my appreciation to Dr. Joan Durso for logistical assistance and the coordination of workshops in which I met Dr. Costanza Bosone who provided support in developing my research proposal.

Abstract

In response to the COVID-19 pandemic, many universities adopted temporary flexible grading policies. At Duke University, these included permitting Satisfactory/Unsatisfactory (S/U) grading to count toward additional degree requirements. Five years later, some pandemic-era policies have been reversed while others remain, raising questions about their lasting influence on student behavior. This thesis evaluates how COVID-19–induced changes to Duke's S/U policy affected voluntary S/U usage using administrative data from the Duke Assessment Office. I model students' likelihood of selecting S/U grading for a course using a logistic mixed-effects model with student-level random intercepts. Significant predictors include period relative to the pandemic, previous-term GPA, course level, course division, course load, and students' academic level. The results provide evidence of persistent post-pandemic changes in voluntary S/U behavior and offer insights to guide future academic policy decisions.

1 Introduction

The COVID-19 pandemic prompted rapid and unprecedented changes in higher education. At Duke University, one of the most visible disruptions was the shift to flexible grading policies. During Spring 2020, all undergraduate courses were graded on a Satisfactory/Unsatisfactory (S/U) basis. Although this universal flexibility was temporary, several related policy changes remain in effect. As the university reconsiders which pandemic-era academic policies should persist (e.g., switching policy defaults so that instructors have to opt-in to allow voluntary S/U) (Halper, 2025), it is essential to understand whether the pandemic experience has produced lasting changes in student grading behavior.

Prior research has examined flexible grading during the height of the pandemic, yet relatively little is known about its long-term effects. In addition, pandemic-era grading occurred alongside substantial disruptions to instruction, assessment, and student well-being, making it challenging to distinguish between temporary behavioral shifts and more enduring ones. This thesis investigates whether student use of voluntary S/U grading at Duke changed in the years following the pandemic and identifies the factors most strongly associated with S/U selection. Using administrative data from the Duke University Assessment Office, I model the probability that a student chooses S/U grading for a particular course, taking into account both student and course-level characteristics. My goal is to provide empirical evidence about the persistence of pandemic-induced behavioral changes and to inform ongoing policy discussions surrounding grading flexibility in the post-pandemic era.

2 Background and Context

2.1 Birth of Traditional A-F Grading

Education is notably missing from the US Constitution, but in the period between 1852 to 1918 all states had passed legislation requiring compulsory education (Diorio, 2023). As a result, K-12 enrollments nearly tripled from 1870 to 1910 (Goldin, 2010). Simultaneously, the Morrill Acts of 1862 and 1890 provided federal land for the establishment of public colleges, opening the doors for access to higher education. Historically, students were evaluated by descriptions and evaluations given by individual teachers. Even entrance exams for college were made by individuals and occurred at individual schools, making results unreliable (Schudson, 1972) (Wechsler, 1977). The massive increase in the number of students required a revolutionary new approach to education that could scale with the increasing demand for education. A standard grading system was needed.

The movement towards the usage of report cards and grades as a success indicator became widespread. Instead, the rise of national examinations such as those by the College Board arose (Valentine, 1987). Now, the shift was towards more consistent report cards across a school or district as a success indicator. Additionally, universities began using academic “credits” to

quantify the amount of work and subsequently the amount of effort a student took on during a given period. Combined with grades, one could compare a student not just within their class but also with students in other classes in the same school. However, grading systems remained variable across the country in regard to what to measure and their frequency (Ashbaugh & Chapman, 1925). Many debates arose ranging from the potential misinterpretation of grades (Bixler, 1936) to the discrepancy between standardized tests and teacher's grades (Hadley, 1954) to the balance of extrinsic and intrinsic motivation (Sumner, 1935). Yet despite the multitude of flaws grading had, the rapidly growing demand for education necessitated a solution.

During the early 1900s, research was conducted to determine the best way to assign grades. Studies such as that by Starch (1913) found that using a 100 percent scale was highly inconsistent across teachers. In response, some suggested a categorical system of "diagnostic letters" to reduce the impact of inconsistency on reported grades (Finkelstein, 1913). By the 1940s, the A-F grading system was adopted by over 80% of U.S. schools, rising in popularity along with the 4.0 scale (Schneider & Hutt, 2013).

2.2 Rise of Pass/Fail

While A-F grading rose to prominence, its core problems remained. Grades were still inaccurate, being assigned differently across instructors, departments, and institutions. Additional studies found poor correlations between college grades and post-educational success (Hoyt, 1966). Concerns about student learning under A-F grading were also raised. As put by Stallings & Leslie (1970),

> The undergraduate perceives grades as that proverbial sword hanging over his head which forces him to study content he otherwise might not study. The power of 'the grade' is strong enough to restrict his studying to material which he anticipates will be on tests (Stallings & Leslie, 1970).

Criticism of traditional A-F grading led to an era of educational innovation. Many schools began experimenting with alternate forms of grading, the most prominent of which was the pass/fail system. Pass/fail grading was not a new idea, with records in American higher education from as early as 1851 (Smallwood, 1935). However, it remained obscure until the 1960s and 1970s. Proponents of pass/fail grading argued that it would foster an intrinsic interest in learning and greater exploration of academic courses. As put by Weller (1983), it was hoped to "free the instructor and the student to communicate on a colleague to colleague basis". By the early 1970s, no penalty grading was present in some capacity in over two-thirds of a sampled 2500 American colleges and universities (Elsner & Brydon, 1974).

Pass/fail grading was not without faults. Multiple studies of the time found that pass/fail grading was often used to concentrate more effort in graded classes to boost or maintain grade

point averages (Quann, 1971) (J. R. Collins & Nickel, 1975). Whether this encouraged exploration outside the major is unclear, with both positive (Sgan, 1969) and negative (Johnson, 1970) (Weems et al., 1971) reports. What was apparent was that students using pass/fail were less engaged in course materials than their graded counterparts. In a study by Karlins et al. (1969), traditionally graded students reported completion of 80% of readings and 85% attendance as opposed to pass/fail students' completion of 61% of readings and 74% attendance. Critics also argued against the binary extremes of pass/fail as well as highlighting administrative challenges regarding the dean's list, calculation of grade point averages, and transfer students. Schools thought that they needed traditional grades to motivate students and that grades convey important information about a student to future employers or higher level educational institutions. Regarding the intent to create bonds between instructors and students, Weller (1983) found that pass/fail grading did not increase faculty evaluation time and institutions were divided on if it had a positive impact on faculty evaluation of students. Nearly 2 to 1 of the pass/fail institutions surveyed believed pass/fail grading did not result in a more positive student perception of grading.

2.3 Decline of Pass/Fail

While research had been conducted on education, the issue of education had largely remained out of the public eye until the 1980s. It was common belief that schools did not matter, and this was given scientific backing by the 1966 Coleman report which found that family background was more influential to student achievement than schools themselves (Coleman et al., 1966). As a result, a relaxed attitude towards academics was commonplace and provided the backdrop for introducing pass/fail.

However, newly emerging research was beginning to suggest that schools did matter. In response to Coleman, the effective schools movement sought to analyze characteristics of schools that correlated with higher academic achievement. Edmonds (1979) expanded upon prior studies such as that of Weber (1971) to analyze practices used by schools with high performing students and outlined characteristics of effective schools. As Edmonds put it, "We can, whenever and wherever we choose, successfully teach all children whose schooling is of interest to us. We already know more than we need to do that. Whether or not we do it must finally depend on how we feel about the fact that we haven't so far." His work was later expanded upon with additional and refined correlates of effective schools by others including Lezotte (1991). Independent research from the UK by Rutter et al. (1979) further strengthened the case for better schools.

In 1983, the National Commission on Excellence in Education published A Nation at Risk. In this monumental report, researchers found consistent declines in high school and college student achievement scores and recommended high school graduation requirements (Gardner & Others, 1983). The report describes the state of American education as "unilateral educational disarmament" and warns of a "rising tide of mediocrity", capturing media attention across the country. Overnight, education became a nonpartisan issue. Pamphlets from the Department

of Education made research more accessible to the public (U. S. Department of Education & Improvement, 1986), the National Board for Professional Teaching Standards was established, exams began to shift away from multiple choice questions, and the first education summit of the nation's governors was held (Ravitch, 1990). As a result of the growing importance of education to the public, schools largely returned to a system of traditional A-F grading. The binary nature of pass/fail grading obscured the student data necessary to measure student achievement and improvement of the education system.

2.4 Rising Educational Attainment

Educational attainment in America rose sharply in the mid to late 1980s as college degrees became increasingly important. Papers from economists found that education was a way to signal and screen for high-ability workers (Stiglitz, 1975). Enticed by the promise of employment, more Americans obtained post-secondary degrees. According to data from the US Department of Education National Center for Education Statistics, the percentage of American adults aged 25 or older who held at least a Bachelor's degree continuously rose from 6.2% in 1950 to 25.6% in 2000 to 37.5% in 2020 (Hanson, 2025). A significant source of this increase was the rise of for-profit "diploma mills." Fueled by the exploitation of financial aid and the political climate of deregulation and privatization, education became an industry. In 1990, there was only a single publicly traded for-profit university; by 2000 there were 40 publicly traded for-profit universities (Beaver, 2017). The Senate Committee on Health Education, Labor and Pensions found that "Between 1998 and 2008, enrollment at for-profit colleges increased 225 percent, compared to 31 percent growth in higher education" (United States Senate: Health & Committee, 2012).

The goal of these for-profit institutions is to cut costs and grow profits. To do so, these institutions prioritized enrollment over teaching. They employed 10 times the recruiters for every career-service employee and hired mostly part-time staff (United States Senate: Health & Committee, n.d.). As a result, the outcomes of students at these institutions have been subpar. Data from the Department of Education suggests that most for-profit career programs fail to benefit students, with 72% of programs having graduates earning less than high school dropouts compared to 32% at public institutions (Education, 2014).

2.5 The Price of Education

In the emerging "credential society", social classes were distinguished by the degree which one held and the prestige associated with that school (R. Collins, 1979). Degrees from elite schools acted as insurance for the future against the rising "fear of falling" of the middle class as household wealth inequalities rose (Ehrenreich, 1989). Backed by impressive scholars and research contributions, admission into these elite colleges has always been challenging. Now, with the oversaturation of degree holders, many would pay whatever price necessary for prestige to ensure financial stability.

At first, institutional rankings originated from athletic college affiliation (Ivy League) and regional primacy (e.g. Duke in the South, USC and Stanford in the West). By the end of the 20th century, third-party ranking systems like that of —U.S. News in 1983—had emerged (*Best College Rankings*, n.d.). An emergent strategy in the battle for the best students has been raising tuition and offering lucrative scholarships to high-achievers.

States have cut funding for public universities since the 1980s. This has only accelerated in the 21st century. In response to the Great Recession (2007-2009) and mandatory spending programs like Medicaid, between 2008 and 2013 the appropriation for the median public university declined by over 20% per full-time student (American Academy of Arts & Sciences, 2014). As a result, many public institutions were forced to increase tuition. A decade after the recession, state funding for higher education has not rebounded in most states (Mitchell et al., 2018).

According to analysis by Banks et al. (2024), annual tuition and fees at private 4-year institutions during the 1979-1980 academic year was $11,357 (adjusted for inflation), compared to the $2,599 (adjusted for inflation) at public institutions. Over time, this gap has widened to a difference of over $20,000 by 2019-2020. By the 2019-2020 academic year, average annual tuition and fees at both public and private 4-year institutions had risen nearly 3 times the cost in 1979-1980, adjusted for inflation. Without adjusting for inflation, the cost of higher education has jumped 10-fold.

2.6 Grade Inflation

As the institutions changed, so did the students. Seeking to distinguish themselves from the increasing number of degree holders and limited by rising tuition, students sought to maximize their grade point averages (GPA) for future profit rather than of pure educational interest. The "entrepreneurial student" shopped "for bargain courses, encouraged by a faculty whose jobs are defined by 'course load', administrators who deal in credit hours as if they were coin, and institutions whose corpus evolves steadily into the corporate" (Haswell, 1999). Yet, not all gains in GPA necessarily match skill.

During the Vietnam War (1955–1975), college enrollment was used to avoid the draft. As a result, failing a student could directly result in their conscription. Evidence has shown an increase in grading leniency due to this policy (Bejar & Blew, 1981) (Birnbaum, 1977). A study by Rojstaczer & Healy (2012) found, "in 1960, as in the 1940s and 1950s, C was the most common grade nationwide; D's and F's accounted for more grades combined than did A." By the end of the Vietnam War, As and Bs made up half to two thirds of grades in American colleges (Davidson, 1975). After the conclusion of the Vietnam War, grades remained a measure of more than a student's academics. Grades were affected by all manner of things from a teacher's concern about student self-esteem, departmental policy to attract students, and the impact of grades during a job search (Schneider & Hutt, 2013).

One of the most predominant reasons for grade inflation was the rise of student evaluation of teaching (SET). SET first began to rise in popularity alongside the Civil Rights movement as a way for students to voice their complaints (Valsan & Sproule, 2008). Under the belief that student evaluations measure teaching effectiveness, administrators realized the opportunity evaluations presented to advertise their programs—with some universities going as far as using evaluations as a component of consideration for promotion, tenure, and resource allocation. By the 1980s, SET became commonplace in American higher education (Centra, 1993) (Wachtel, 1998). Yet, "the typical SET questionnaire treats the student as a customer and measures the satisfaction of the student with his or her professor, and not learning" (Crumbley et al., 2010). Multiple studies have found significant, positive correlations between student evaluations and student grades (Langbein, 2008) (Ellis et al., 2003). On the other hand, SET rankings are not significantly correlated with actual student learning (Uttl et al., 2017). An article by Neath (1996) titled "How to Improve Your Teaching Evaluations Without Improving Your Teaching" even goes so far as to suggest multiple methods such as getting evaluated before exams and grading leniently. Simultaneously, efforts to cut costs and increase profit margins resulted in a rise of nontenured, adjunct faculty (Bettinger & Long, 2010). These educators' careers depended significantly on SET rankings. As such, over time, professors became increasingly aware of the implications SET rankings could have on their careers.

Grade inflation is unevenly applied across institutions and subjects. Average student GPAs in private schools have historically been higher than their public counterparts. This is in part due to the selection of high-achieving students, but pre-college performance does not completely explain the difference. A 2010 study by Rojstaczer and Healy analyzed patterns within their database of over 160 colleges and universities from 1920 to 2006. They found that, on average, private school students were graded 0.1 to 0.2 higher on a 4.0 scale for a given caliber of student (measured with SAT scores or a selectivity measure). When looking at grading across divisions, they found that on average science departments grade 0.4 lower on a 4.0 scale than humanities departments and 0.2 lower than social science departments. Evidently, both the institution and the division of a student's courses are correlated with a student's GPA.

Despite the increase in proportion of A's, this does not seem to reflect an increasing caliber of student. According to the 2019 National Assessment of Educational Progress High School Transcript Study, the average GPA has increased from 3.00 in 2009 to 3.11 in 2019; over the same time period Grade 12 assessment scores decreased in mathematics and did not significantly change in the sciences (*2019 NAEP High School Transcript Study (HSTS) Results*, n.d.). Additionally, it is still debated as to how good of a predictor high school GPA is compared to standardized tests like the ACT and SAT. In a study comparing the predictive power of high school GPA against composite ACT scores, Noble & Sawyer (2004) found that across all levels of achievement, ACT scores provide greater differentiation than high school GPAs on success in the first year in college. In particular, "at 93 percent of the institutions, a student with a 4.00 high school GPA had less than a 0.50 probability of earning a 3.75 or higher first-year GPA" in higher education and "in some cases, HSGPA values less than 3.00 provided little differentiation in terms of students' chances of achieving different first-year GPAs" (Noble & Sawyer, 2004). On the other hand, some studies show that ACT scores and high school

GPA are both valid predictors of first year performance (Westrick et al., 2015). Overall, there is insufficient evidence to suggest that the quality of students has risen, despite significant increases in GPA. GPA is no longer an effective tool for differentiating skill.

2.7 Falling Confidence in Higher Education

Grade inflation devalues education. The weight of a 4.0 GPA no longer carries the weight it once did. For colleges, participation in grade inflation lessens rigor, lowers quality of education, and degrades reputation (W. Chan et al., 2007) (Ehlers & Schwager, 2016). For students, they are deprived of feedback and left unmotivated by lack of recognition of exceptional effort (O'Halloran & Gordon, 2014). For society, grade inflation means graduates are unprepared for the workforce without the skills, dedication, knowledge, and work ethic desired by employers (Iris Franz, 2010) (Love & Kotchen, 2010) (Yang & Yip, 2003).

With more degree holders and higher GPAs, degrees are no longer a sufficient edge over other job searchers to obtain employment. Simultaneously, the rate of tuition increase has outpaced wages. This discrepancy has not gone unnoticed. Multiple news sources have published articles with headlines such as "Price Of College Increasing Almost 8 Times Faster Than Wages" in Forbes (Maldonado, 2018). This crisis has been expedited by rising housing costs and other costs of living. In order to afford degrees, federal student loan debt increased by over seven-fold between 1995 and 2017 (Burk & Perry, 2020).

Increasingly, the American public has been losing trust in higher education. According to a 2024 Gallup poll, reported confidence in higher education has fallen since 2015 from over 65% down to 36% in 2024 (*U.S. Confidence in Higher Education Now Closely Divided*, n.d.). Meanwhile, the percentage of people reporting very little/no confidence has tripled from approximately 10% to 32%. Additionally, the gap in unemployment rates of Americans aged 25 and up by educational attainment has shrunk. What used to be a 5% difference in unemployment rates between those without a high school diploma and bachelor degree holders in 2005 is nearly halved in 2025 (*Unemployment Rates for People 25 Years and Older by Educational Attainment*, n.d.). However, the relative difference in median income of high school graduates and bachelor degree holders has stayed roughly the same since 2004, adjusted for inflation (Scherer & King, 2025).

2.8 Flexible Grading in the 21st Century

In the early 21st century, research and usage of pass/fail grading remained largely obscure. The most prominent use of pass/fail grading was found in medical schools. Doctors are expected to be lifelong learners, staying up to date with the newest techniques, treatments, and health problems. To do so, the character of a doctor must be taken into account, particularly their intrinsic desire to learn. Additionally, higher than average rates of stress and burnout had been reported in medical students and the negative effects of distress have been well studied

(Shapiro et al., 2000) (Dyrbye et al., 2005). As a result, a wave of medical reforms were made including: resident duty restrictions, self-development groups, and pass/fail grading systems.

In a 2011 review of pass/fail and well-being literature (1980-2010) in medical schools, it was found that all (four) of the papers reported improvement in some measure of well-being (stress, anxiety, depression, self-control, good health, level of satisfaction, group cohesion, and amount of free time) (Spring et al., 2011). Student satisfaction was measured and found to have increased in two of the papers (Bloodgood et al., 2009) (Robins et al., 1995). However, there were discrepancies in the long-term effect of pass/fail, with some claiming continued effect after the first semester while others found a return to typical levels in later semesters.

Spring et al. (2011) also reviewed an additional five papers (9 total) on pass/fail and academic outcomes (GPA, scores, residency attainment and performance). Grades were not found to be significantly different between pass/fail and tiered grading cohorts. Pass/fail cohort average was significantly higher than the pass/fail cut-off, signaling more effort than the bare minimum (Robins et al., 1995). The pass/fail system was not found to adversely affect academic performance. However, acceptance into desired residency programs may be negatively impacted by pass/fail. In terms of residency attainment, roughly 73% of directors claimed they did not give preference to tier-grading schools and 33% of programs who filled all spots preferred students from tier-graded schools (Tardiff, 1980). Surveys of students and directors also showed majority belief that pass/fail evaluation hindered the ability to compete for residency (Dederichs et al., 2020) (Provan & Cuttress, 1995). Although, other studies have shown that some students believe grades are already arbitrary while others prefer grading for motivation (Dederichs et al., 2020). However, it appears that the actual impact of grading tiers may depend on the institution giving the grades. When program directors were asked to compare Stanford's pass/fail classes with their own classes, only 3% of the Stanford graduates were judged as "poor" compared to their peer group (Vosti & Jacobs, 1999). There was also evidence that pass/fail grading reduced competition and external motivation for grades without decreasing the amount of time students spent studying, defying expectations of increased laziness (Jessee & Simon, 1971).

Research supporting pass/fail and flexible grading continued to be published in the late 2010s. For instance, a decade-long longitudinal study of medical student well-being found an 85% decrease in depression rate and a 75% decrease in anxiety in first-years when switching from a four-tier to two-tier grading system and restructuring their curriculum, among other changes (Stuart, 2019). Additional studies on the differences between pass/fail and graded students found no consistent difference between student cohorts (Ange et al., 2018). Interestingly, some have also suggested the usage of pass/fail as a method of combating grade inflation (Blum, 2017) while others argue pass/fail is a cause of grade inflation (Goldman, 1985).

In summary, there has been an increased interest in pass/fail arising from a prevalent attitude that educational reforms are necessary. Researchers have found evidence that flexible grading systems reduce student distress, support collaboration, and encourage intrinsic learning without having a substantial impact on academic performance and test scores (White & Fantone, 2010). However, there is also evidence that suggests that a two-tiered system makes

it near impossible to distinguish between satisfactory and truly exceptional students, harming future career prospects as well as disincentivizing some students from putting forth their best effort and persevering through challenges. There is no definitive consensus in the literature on whether or not pass/fail is a better system than traditional A-F tiered grading.

2.9 Education During the COVID-19 Pandemic

Disaster preparedness, response, and relief is an important role that education fulfills. Schools help educate the public about how to prepare and act during disasters, and the return to school can serve to ease stress with its familiarity (Mutch, 2014). There is a plenitude of literature on crises such as school shootings (Beland & Kim, 2016), political conflicts (Brück et al., 2019) (Sullivan & Simonson, 2016), and natural disasters (Shaw et al., 1996) (Sacerdote, 2012) (Devaney et al., 2009). However, the COVID-19 pandemic was at an abrupt, unprecedented scale, affecting the daily lives of nearly all communities.

On March 11, 2020 the World Health Organization declared a global pandemic. Policies such as travel restrictions, telehealth, social distancing, stay-at-home orders, and screening were implemented. In education, the response included remote learning, flexible deadlines, alternate assessment strategies, and relaxed grading policies. There is a growing body of research on the immediate and long-term effects of the pandemic on education. In the aftermath of widespread school closures in Spring 2020, there is evidence of a negative effect of school closures on student achievement (Hammerstein et al., 2021). In the three years spanning the pandemic (2020-2023), test scores were observed to have fallen compared to pre-pandemic levels and achievement gaps were amplified (Kuhfeld et al., 2022). Numerous other studies support the disproportionate impact of the pandemic on vulnerable groups (Perry et al., 2021) (Cross et al., 2022) (Rodríguez-Planas, 2022). Yet, grade inflation during this period exceeded trends and still persists in some academic divisions (Kuperman et al., 2025). In addition to cases of COVID-19, students' physical well-being trended down as there was less engagement in physical activity (Neville et al., 2022). Similarly, mental well-being also declined with an increased worldwide prevalence of depression and anxiety (Santomauro et al., 2021). A plethora of research supports the idea that the pandemic exacerbated the pre-existing youth mental health crisis in facets from eating disorders to peer connectedness to substance use (Thorisdottir et al., 2021) (Prichett et al., 2024) (Widnall et al., 2022) (Agostino et al., 2021).

2.10 Intersection Between Flexible Grading and the Pandemic

The focus of this thesis is the impact of the pandemic on flexible grading policy. In the wake of school closures and remote learning, state Department of Education guidance often suggested or even required alternative grading (Townsley & Kunnath, 2022). A review paper by C. K. Y. Chan (2023) on COVID-19 academic changes in higher education identified binary grading as one of their five key themes. At least 194 American universities implemented pass/fail policies during the Spring 2020 semester (Gibbs, 2020). However, there is a dearth of research

specifically evaluating the impact of COVID-19-induced flexible grading. One study using data from Queens College suggested flexible grading policies helped reduce negative impacts of the pandemic, particularly for lower income students (Rodríguez-Planas, 2022). Another study using data during the first three semesters of the pandemic from a historically black college and university in North Carolina found that: utilization of flexible grading varied significantly between subjects; flexible grading was less likely to be used in general education courses; flexible grading use varies across socio-economic groups, first-year students were more likely to choose flexible grading; and STEM students were less likely to use flexible grading (Mostafa et al., 2023). When looking at some of the most popular course sequences, they noticed a mix between students who benefited and those who were disadvantaged in the subsequent course. Perhaps related to this is a finding that in a study of two Canadian universities from the 2018-2019 to 2022-2023 academic year, in some subject areas GPAs have returned to pre-pandemic levels while in others GPAs remain inflated (Kuperman et al., 2025). While there is no question flexible grading changed during the pandemic, its long-term effects are unknown.

While a couple of studies do examine the interaction between pandemic policies and flexible grading on student outcomes, their timeframes struggle to capture post-pandemic changes as the pandemic was not officially declared over until May 2023. While changes in behavior occurred at the height of the pandemic, it is unclear whether they persisted as education returned to pre-pandemic levels or if they will stabilize at a new norm. Additionally, both Rodríguez-Planas (2022) and Mostafa et al. (2023) used data from only moderately selective, affordable, and large public universities. There is a lack of research on the impact of pandemic policies on flexible grading on highly selective, prestigious, expensive, or private universities.

3 Institution of Study

Duke is a private, non-profit university located in Durham, NC. Duke is a highly prestigious and selective university with an overall acceptance rate of 4.8% for the most recent Class of 2029 (*Duke Welcomes The Newest Members of the Class of 2029*, 2025). Duke University has roots beginning in the 19th century, but was officially established in 1924 (*Duke University*, 2020). Over the past 100 years, grading policies have shifted with the times. I am uncertain when letter grades began usage at Duke University, but the language referring to each letter fluctuated in the Undergraduate Instruction Bulletins from the mid-1950s to the late 1960s (Registrar, n.d.). For instance, in the 1955-1956 academic year, a C was described as "medium" and a D as "passing." Yet in the following year, Cs were "average" and Ds were "inferior." Throughout the 1960s, Ds denoted "low pass."

In addition to adjustments in association of letter grades to adjectives, more fundamental policy changes began to take root. In 1960, the first mentions of auditing can be found in the undergraduate bulletins. When a student audits a course, it shows up on their transcript but no grade or credit is given. They have the option to complete assignments and take exams. In 1966, + and - grades were introduced and associated with numeric "quality points" (equivalent

to GPA thresholds). In the same academic year, historical bulletins show the introduction of pass/fail grading. Since then, grading policies have remained relatively consistent except for a shift from pass/fail to satisfactory/unsatisfactory—allows students to S/U before major declaration (P/F and S/U thresholds are identical).

At Duke University, the cutoff for receiving an S is a C-, so all grades D+, D, D-, and F correspond to a U (*Courses*, n.d.). S/U grading can appear in two contexts: involuntary or voluntary. The type of S/U allowed is dictated by the course instructor; if a course is S/U required, all students—regardless of status—will be graded S/U. Voluntary S/Us must be requested before the withdrawal deadline and approval is required for the student's academic Dean. In order to apply for voluntary S/U, the student must also be enrolled in "at least 4 x 1.0 credit" courses. The decision to switch to voluntary S/U also depends on whether or not the student's academic programs allow S/U courses to count towards their degree and how many of the four voluntary S/U credits permitted to count towards graduation requirements are left. However, there are notable pandemic-related exceptions.

In Spring 2020, COVID-19 cases and concerns prevented a return to the classroom after spring break. In an email to faculty on March 18, the university announced a shift of all courses to S/U grading with the option of students opting back in to receiving a letter grade (*Duke Coronavirus Response*, n.d.). For Fall 2020, departments were allowed to convert any of its 199 or below level courses to a mandatory S/U grading basis. Throughout the 2021-2022 academic year, periods where in-person classes were prohibited persisted along with routine COVID-19 testing policies. Fall 2022 was the first full semester where classes were not interrupted by periods of remote learning. Masking, gathering, and other pandemic-era policies had also been systematically loosened. In light of this historical context, I define the pandemic as ranging from Spring 2020 through Spring 2022. The periods before and after will be referred to as pre-pandemic and post-pandemic, respectively.

All students in my dataset fall under Trinity College's Curriculum 2000. To provide a brief overview, the general education requirement mandates that students take a minimum of two courses that fall under each of several categories. For instance, the five Areas of Knowledge are: Arts, Literature, and Performance (ALP), Civilizations (CZ), Natural Sciences (NS), Quantitative Studies (QS), and Social Sciences (SS) (*Curriculum*, n.d.). Undergraduate students matriculating in Fall 2025 and onward (outside the scope of my analysis) follow a new Arts & Sciences Curriculum (*Curriculum for Students Enrolling in Fall 2025*, n.d.). Prior to Spring 2020, S/U grading was only permitted to count towards total graduation credits. After Spring 2020, S/U grading has additionally been permitted to count towards general education requirements.

4 Data

4.1 Note on Ethics and Data Inaccuracies

All data used in this thesis are sourced from the Duke University Assessment Office through Jennifer Hill. Rows represent student-course pairs. In order to protect student privacy, data was masked and no socio-demographic data was provided. Furthermore, all data remains on in-office devices and I will no longer have access upon completion of my thesis. Please contact Jennifer Hill if you have questions/concerns about my base dataset.

While these data come from official records, how data are recorded is not always straightforward. Notably, it remains unclear to me what the exact algorithm for calculating a student's academic level is. I experimented with matching a student's current semester number with their reported academic level. While there are some discrepancies, I choose to believe the algorithm for academic level must include other factors such as progress towards degree completion. I also noted some unintuitive values for total enrollment in which I occasionally found classes claiming zero enrollment, yet that student still received a grade. After discussion with the Assessment Office, we believe it may be due to enrollment in a different section of that course being mistakenly used instead of totaling enrollment across sections. For instance, a course may be cross-listed in department A and department B but instruction is identical in both sections. While other potential accuracy issues may be present in my data, they should have a relatively low impact on my overall analysis.

4.2 Data Cleaning

The original dataset I was provided included records ranging from Fall 2006 to Spring 2025. However, S/U grading only began in Fall 2012. Hence, I drop all observations prior to Fall 2012. Since my interest is in the impact of COVID-19 policy changes within the Trinity School of Arts and Sciences, only students who graduated from or intend to graduate from the Trinity School of Arts and Sciences are included in this thesis. There may be a number of current students who later decide to transfer to Trinity or transfer to Pratt, but I do not expect this number to be significant enough to be of concern. I also exclude any students who transferred to Duke University from another institution. Additionally, my focus is on the "typical" student, whom I define as a student who graduates in exactly eight Fall/Spring semesters. For students who matriculated in 2022 onward and hence have not had the opportunity to enroll in eight semesters, I assume all students will graduate in eight semesters (although this may not be the case).

In addition to dropping the small fraction of courses that were unable to be matched (missing values for division, catalog level, and other course-specific attributes), I dropped courses belonging to the division of "Other." Broadly speaking, "Other" includes courses such as those for ROTC, Robertson Scholars, music lessons, and house courses. These courses are typically

not taken purely for an "academic" purpose, and hence I do not consider them as qualifying a student to be in a true academic overload. In my opinion, half-credit music, dance, and PE courses also fall into a similar category as more of an extracurricular activity. I refrained from dropping all less than full credit courses from GPA and course load calculations due to the existence of courses which give credit to labs and other mandatory courses, such as the half-credit STA 211 which was required for the major in Statistical Science.

For modeling, my dataset has also been filtered to only contain courses taught online or at Duke's campus at Durham during either the Fall or Spring semester. This removes the impact of differences in environment, pace of instruction, grading institution, and other factors that confound these semesters and courses. Online courses are kept due to their prevalence during COVID-19 and their persistence for niche courses such as those in the Cherokee Language Program. For courses where location was not recorded (17.6% of student-course pairs), I assume they were taken at Durham. This may result in an underestimate of students who study away.

My goal is to model student choice of S/U over A-F grading. Hence, I drop all withdrawals, audits, and other non-credit courses. Additionally, Trinity College states that students "must be in a normal course load (at least 4 x 1.0 credit classes) to request to change a class to S/U" and that "you may only request a Voluntary S/U for a single 1.0 credit course" (*Courses*, n.d.). Hence, while I consider all "academic" courses in GPA and course load calculations, only single full credit courses are eligible to be included in my model. I use actual course loads to drop student-semester pairs with less than 4.0 credits. I did not verify that students have at least four full credit courses.

A major roadblock I faced is the possibility of S/U only and graded only courses in which students have no say in what grading method is used. Again, I lack this data and once again make assumptions. For classes that have 99% of all students receiving non-letter grades (i.e. S, U, withdrawal), I assume they are S/U required courses like ECON 101. I believe this is reasonable in most cases as it is unlikely for all students to make the same decision, unless the class size is small. These mandatory courses are dropped from my model. However, I do not assume the converse is true. If all students in a course receive an A-F (or withdrawal, etc.), I choose not to assume the course is graded only. My belief is that in the majority of cases, students would rather take a course on a graded basis as Trinity only allows 4.0 credits taken S/U to count towards graduation requirements. However, I do acknowledge that this means it is likely that my model will underestimate how likely a student is to take a course S/U in cases where the instructor prohibits S/U.

4.3 Feature Engineering

I engineered several variables as I describe below:

- `timeperiod` is defined relative to COVID-19 policies with Fall 2012-Spring 2020 labeled as pre-pandemic, Spring 2020-Spring 2022 as during pandemic, and Fall 2022-Spring 2025 as post-pandemic
- `is_art_humanity`, `is_natural_sci`, `is_social_sci` are each binary indicators for whether or not a student has a major that falls under the division
- `prev_semGPA` is the weighted GPA from the semester prior (or 4.0 if this is their first semester), only A-F grades count towards GPA
- `num_plans` is the number of majors, minors, and certificates a student holds
- `took_summer_courses` is a binary indicator if a student has ever taken summer courses
- `studied_away` is a binary indicator if a student has ever taken a course somewhere other than in Durham or online
- `actual_units` is the true number of academic credits a student is taking in a semester
- `actual_load` is a categorical factor with levels of underload (term_units < 4.0), normal (term_units $= 4.0$), and overload (term_units > 4.0)
- `term_units` is the number of academic credits (excludes "Other" courses, P.E., etc.) a student is taking in a semester
- `load_status` is a categorical factor based on academic credits with levels underload (term_units < 4.0), normal (term_units $= 4.0$), and overload (term_units > 4.0)
- `num_overloads` is the number of academic overloads a student has taken so far

4.4 Exploratory Data Analysis

Before fitting any models, I conducted a brief exploratory analysis. The dataset used for Figures 1-4 is larger than my model dataset. The dataset they use has dropped study away and summer courses, as well as only retaining Trinity undergraduates who graduated in eight semesters or non-transfer students who matriculated in 2022 and beyond. Only Figure 5 was created using the modeling dataset from the process described in the Data Cleaning section, above.

In Figure 1, I plot changes in the proportion of student-course pairs which have the final grade of either "S" or "U". The results verify my expectation of a massive surge in S/U grades during the Spring 2020 semester in which all courses could be taken S/U. Additionally, it also reveals a large discrepancy between Fall 2012-Fall 2019 with the total proportion of grades being taken S/U less than 2.5% of the time, compared to averages around 5% in Fall 2020 through Spring

2025. A very similar plot (not shown) was drawn with counts on the y-axis as the number of enrolled students and the approximate number of courses taken per term stays consistent across all years.

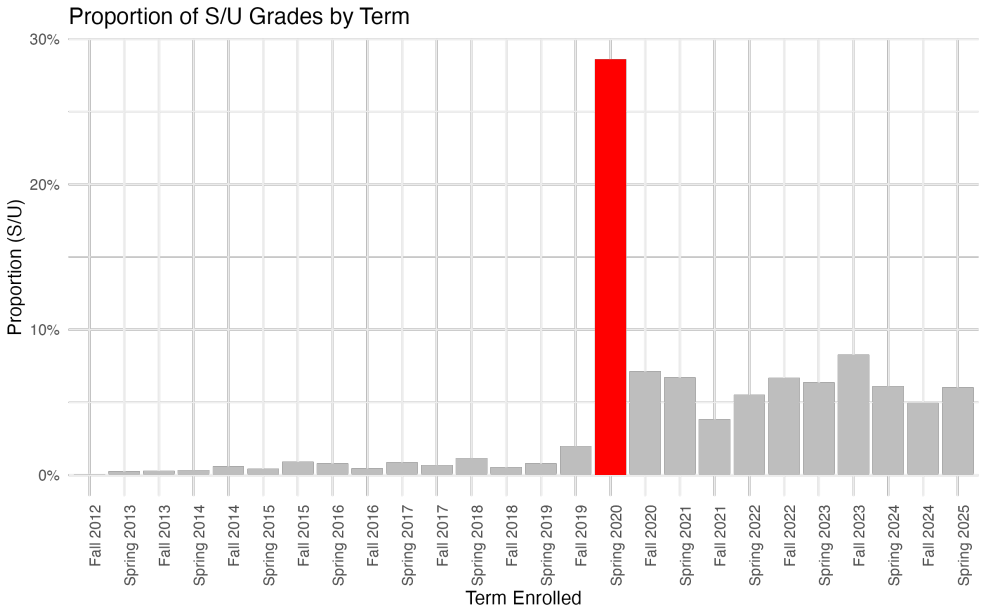

Figure 1: Bar chart of the proportion of all grades given to Trinity College undergraduate students taught at Durham or online in Spring and Fall terms from Fall 2012 to Spring 2025. The prominent, red peak in Spring 2020 indicates the impact of pandemic policies during that term. Prior to the pandemic, there seems to be a slight increasing trend in S/Us. In contrast, terms following Spring 2020 are relatively stable at a higher proportion of S/U grades with no increasing or decreasing trends. Note that changes in S/Us may be related to the number of courses with mandatory S/U grading.

To put the shift in S/U grades into context with all grades, Figure 2 uses the same data but shows relative proportions of grade categories. Over time, it's clear that the proportion of A's given has steadily increased since Fall 2012, with a dip in Spring 2020 likely due to pandemic impacts as S's rose sharply. The proportion of S's seemingly surpassed D's within the first five years of its introduction. From the pandemic onward, the proportion of S's have also overtaken the proportion of C's. From the declining trend in B's, it is possible that—without policy changes—S's may even surpass B's in the coming academic terms. Additionally, the prominence of non-credit grades over D's seems to signal the lack of use a D has in the modern academic system. While beyond the scope of this thesis, it could be hypothesized that students may be opting for non-credit grades to avoid drops in GPA. What is striking is that in the 2024-2025 academic year, for the first time (as shown in Figure 2), A's made up over 80% of all grades, providing evidence towards continued grade inflation in Duke University's Trinity

College of Arts & Sciences.

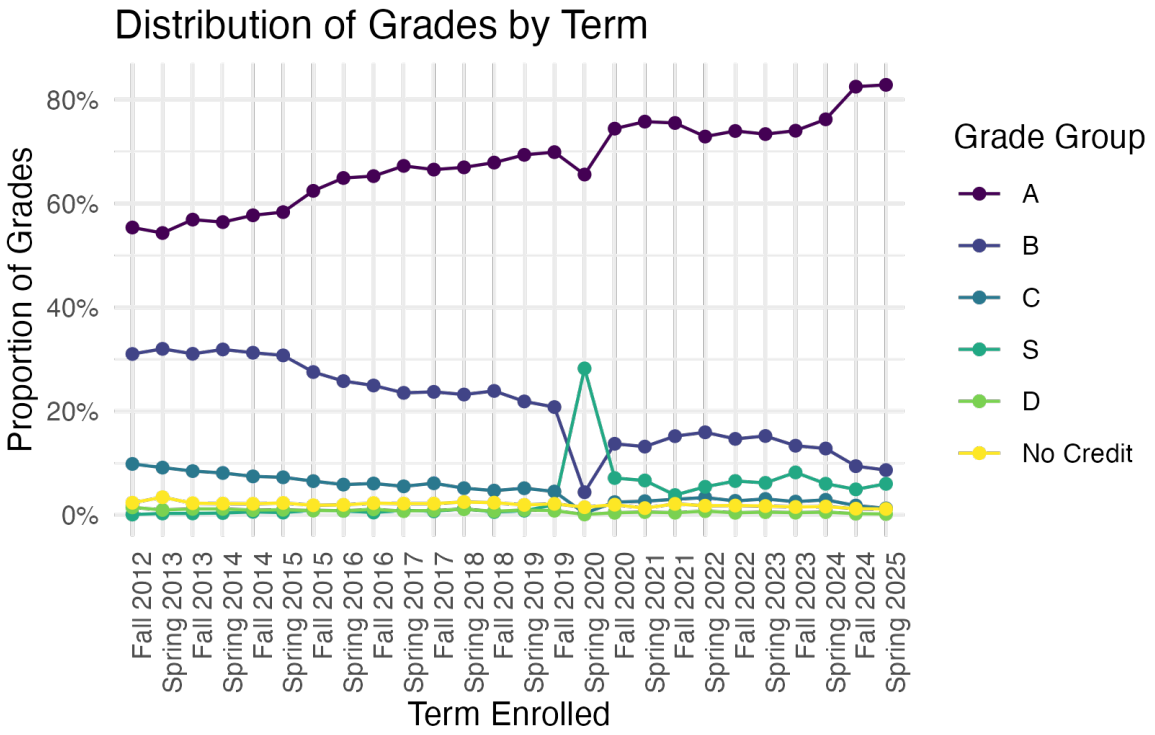

Figure 2: Line plot of relative proportions of final grades in each of 6 categories. Plus and minute (+/-) letter grades are together for ease of interpretation, although +/- grades are associated with different GPA points (with the exception of A+ and A). "No Credit" refers to a variety of grades including failing grades (F and U), withdrawals (W), incomplete (I), audit (AD), and other miscellaneous grades that do not impact a student's GPA.

For more clarity on grade inflation at Duke University, I plot median and quantiles of weighted semester GPAs in my dataset. The plot is also supplemented by historic Dean's List (a distinction awarded for students in top class rank percentiles) cutoffs from the Office of the University Registrar *Academic Honors and Recognition* (n.d.). From Figure 3, there is a clear trend of increasing weighted GPAs throughout the period from Fall 2012 to Spring 2025. Understandably, there is a spike in GPAs during the pandemic. Yet, even with a return to normalcy in Fall 2022, GPAs continue to steadily rise. In the 2024-2025 academic year, median weighted GPAs have reached the maximum of 4.0. Perfection is now the norm.

While the previous figures have been influential in showing trends in final grades across terms, they are somewhat uninformative in the sense that they have no distinction between voluntary and mandatory S/U grading. Figure 4 presents a summary of the proportion of all courses taken by Trinity undergraduates from Fall 2012 through Spring 2025 by their perceived grading

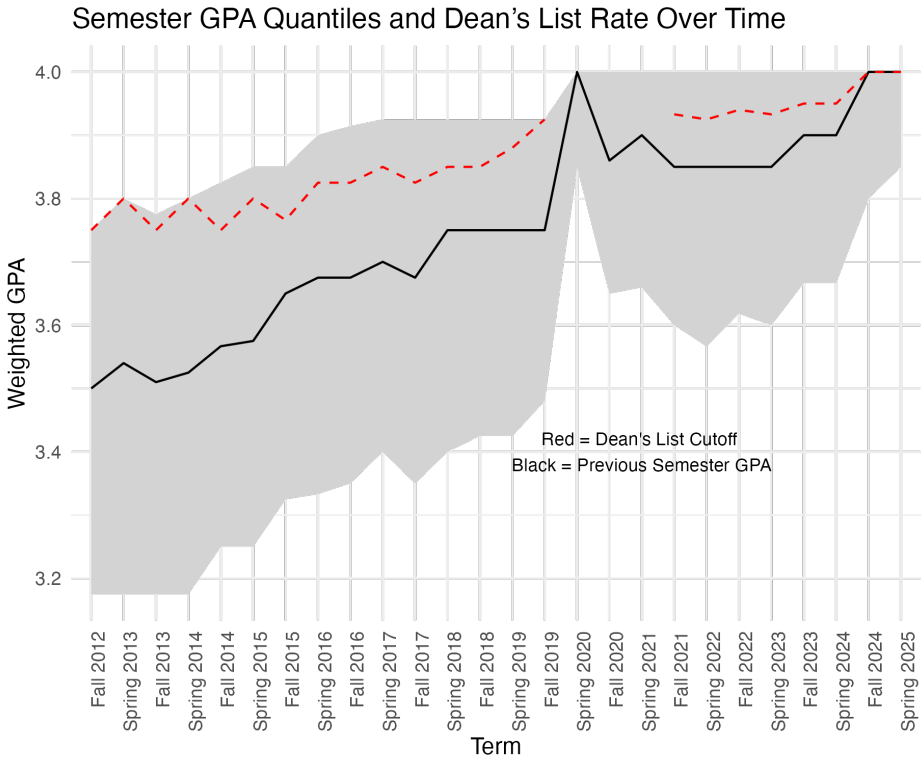

Figure 3: Plot displaying trends in Trinity undergraduate students' weighted GPAs from Fall 2012 through Spring 2025. In black is the median semester weighted GPA from my dataset. The gray bands represent the 25th and 75th percentiles of weighted GPAs. Missing values for first semester students are not included in this figure. Overlaid in red are the historic cutoffs for the dean's list in Trinity. The dean's list was paused in Spring 2020, Fall 2020, and Spring 2021 due to pandemic-era policies.

status. As a reminder, I do not have information on whether a course is officially classified as "Graded Only" or "S/U required." Instead, if no students received grades A-F I classify the course as "S/U Required." If no students received grades S or U, I classify the course as "All Graded." Overall, there seems to be a growing proportion of courses in which some students receive S/U grades and a corresponding decline in the proportion of courses where only A-F grades are given. This is indicative of a rising number of courses in which students choose to voluntarily opt for S/U grading (the number of courses offered has remained relatively stable). However, this plot does not indicate or weight by the size of enrollment.

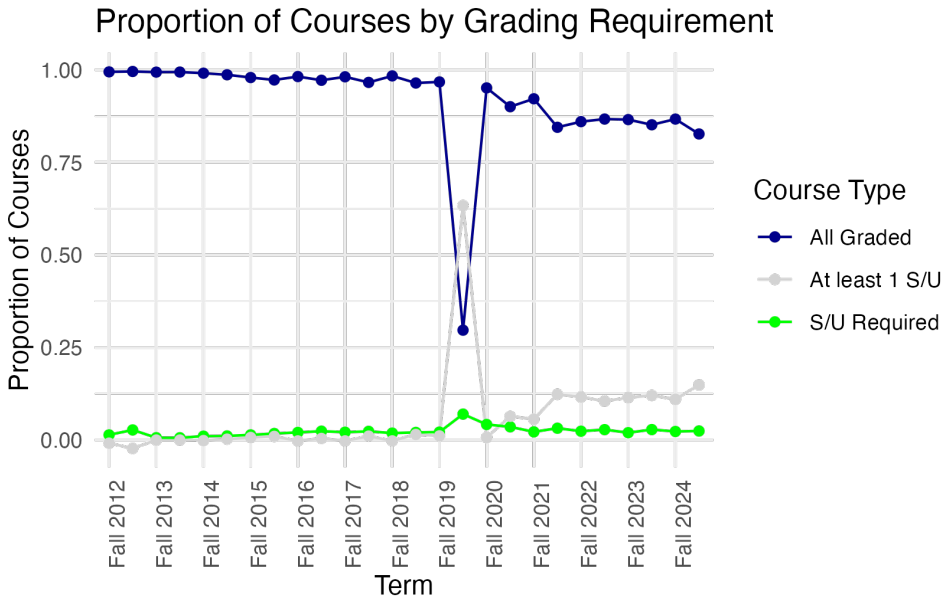

Figure 4: Line plot showing relative proportions of courses that are taken on varying grading bases. Rather than being based on if all grades are S/U, it instead checks if any of the grades are A-F in order to account for alternative grades such as W and AU. Overall, there is a declining trend of courses with no S/U grades and an increase in courses with S/U grades.

Figure 5 seeks to quantify the effect of S/U courses at the student level. For this figure, courses categorized to be S/U required have already been dropped and the data matches that used by my models. It's important to note that dashed lines represent current statuses and the actual number of S/U courses taken may be higher as those students have more semesters left before graduation. While all students plotted have or are projected to take eight semesters, they may have taken more or less courses. For students who matriculated in 2012 through 2015, there was little change in the number of S/U courses. However, for those who matriculated in 2016 through 2019, there was a steady increase in the number of voluntary S/Us per student. These classes of students are precisely the ones who were enrolled during Spring 2020. The first class of students who matriculated after Spring 2020 (student group 2020) show a lower

proportion of voluntary S/Us than those who experienced Spring 2020. However, the student group matriculating in 2021 on average have taken more voluntary S/Us than the student group for 2020, despite even less time experiencing pandemic era policies. For students who have yet to graduate, they are still projected to have taken more voluntary S/Us than their 2012-2015 counterparts.

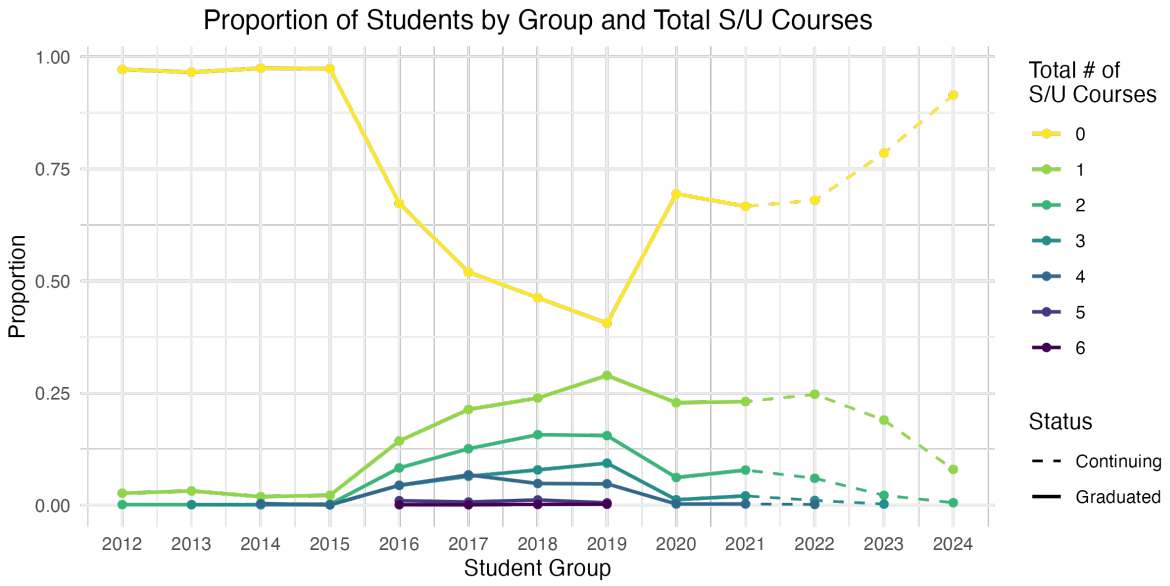

Figure 5: Line plot colored by total number of voluntary S/Us taken by students who matriculated in the Fall of years 2012 through 2024. The y-axis contextualizes for what proportion of students in each group that number of voluntary S/Us is achieved. Solid lines represent final values for classes who have already graded. Dashed lines represent current values for classes that have not yet graduated, meaning they are conservative underestimates of the total number of voluntary S/U courses.

5 Methodology

All analysis in this thesis is performed using R. Following Mostafa et al. (2023), I use logistic regression to model whether or not the student selected a flexible grading option for a given course.

5.1 Data Imputation

Since post-pandemic policy allows students to use S/U for general education requirements, I believe that changes in S/U will likely be reflected in general education courses. Curriculum

2000, under which all students in my data fall, general education requirements mandate that Trinity students to take two credits in each of five Areas of Knowledge as well as two credits in each of five Modes of Inquiry. Unfortunately, I was unable to obtain these course codes as part of my dataset.

As a proxy for determining general education courses, I attempt to classify students by their academic plans under the assumption that most of the courses taken outside of the division(s) of their major(s) are taken for general education requirements. Trinity students are permitted to have up to three academic plans (majors, minors, and certificates) of which only two can be majors. Since minors and certificates tend to only require 5 or 6 courses and tend to be added towards the end, I do not classify students as belonging to divisions based on their minors and certificates.

Due to privacy concerns and the potential to identify individuals given full records, course names, numbers, and departments were omitted. Instead, I was given catalog level (i.e. 100-199, 200-299) and course division (Engineering, Arts & Humanities, Social Sciences, Natural Sciences, and Writing). In order to match majors with course divisions, I manually assigned each major with the appropriate division based on classifications at https://trinity.duke.edu/. While most majors were relatively straightforward, I was unable to classify Program II and unlabeled interdepartmental (IDM) majors. These students are excluded from my model. The spreadsheet I used for classification of majors can be found in my GitHub repo.

In addition to ambiguous majors, there are also a significant number of unknown majors in my data. Student academic plans in my dataset come from post-graduation records. While some missing values are expected for the small percentage of students that do not graduate (dropped from my model), there is a more substantial problem: I lack any graduated students for the Class of 2026 (matriculated in Fall 2022) onward. Student academic plans are missing not at random, and if I was not to address this my analysis would be biased. Additionally, note that this is precisely the classes of students who fall after my pandemic time period (Spring 2020 – Summer 2 2022). Since Duke undergraduates are required to declare their major(s) by the end of sophomore year and due to how little data there will be for first-years, I cannot possibly impute majors from current first-years and sophomores (Class of 2028 and 2029). However, I do attempt to impute major(s) for the Class of 2026 and 2027.

I used labeled data to inform my imputation, assuming that there has been no significant change in the distribution of courses a student takes in their major(s). Initial attempts included experimentation to determine the best threshold for number of courses and major divisions. However, using number of courses taken does not extend well to students who have not completed a full eight semesters. In fact, it is plausible that some students remain undecided in their first several semesters, some students space out their requirements across all four years, and some students focus on their major and only later are reminded of general education requirements.

My current approach is to use the proportion of 200 and above courses a student has taken per division to assign major divisions. With a 93%, 87%, and 87% accuracy for major divisions,

the data suggested a threshold of 0.5, 0.4, 0.3 for Arts & Humanities, Natural Sciences, and Social Sciences respectively (Figure 6). The high threshold for Arts & Humanities logically makes sense; the more classes a student takes in a division, the more likely they are to be majoring in that division. However, the threshold for Social Sciences seems abnormally low as it insinuates a student who has a major in Social Sciences only takes roughly a third of their courses in Social Sciences. Looking at the confusion metrics for each proportion threshold in Figure 7, it becomes evident that the increase in false positives with higher thresholds offsets the increase in true positives. It should also be taken into consideration that 1) many students pursue a double major, 2) students may have minors or certificates in other divisions, 3) students have a variety of academic interests, 4) courses may be cross-listed across divisions, and 5) categorization of majors is somewhat debatable. To provide an example of 5, as defined by Trinity College, African & African American Studies is classified as a Social Science while Asian & Middle Eastern Studies is in Arts & Humanities.

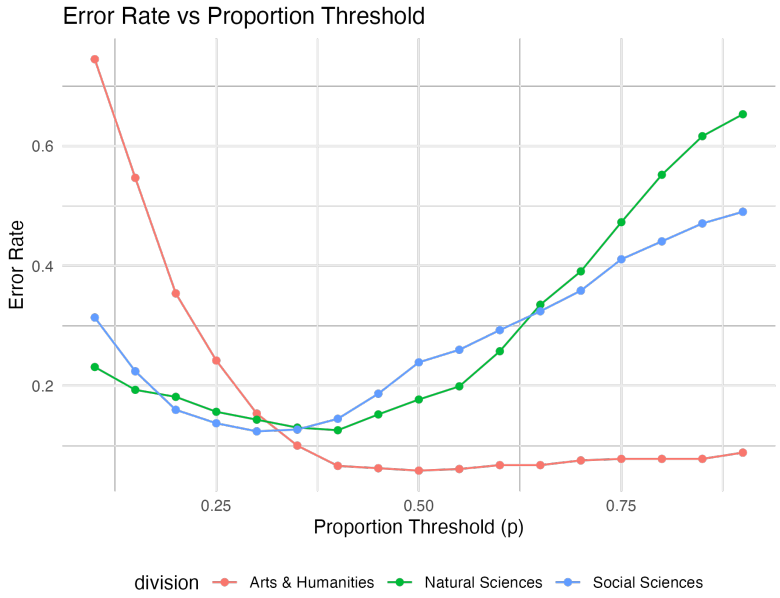

Figure 6: Test results for different thresholds for major division imputation. For the Arts & Humanities, students appear to solely take courses in their home division with higher cutoffs performing similarly when compared to lower cutoffs. However, for Natural Sciences and Social Sciences, higher thresholds resulted in more errors.

To ensure that this imputation strategy is not dependent on when a student is enrolled, I also tested this imputation strategy by class year. Results in Figure 8 show slight variations in the best threshold and error rates across class years. However, there does not appear to be a significant correlation between student group and thresholds for imputation. Hence, I deem that my approach is appropriate to be applied to the subsequent classes of students in which I am imputing.

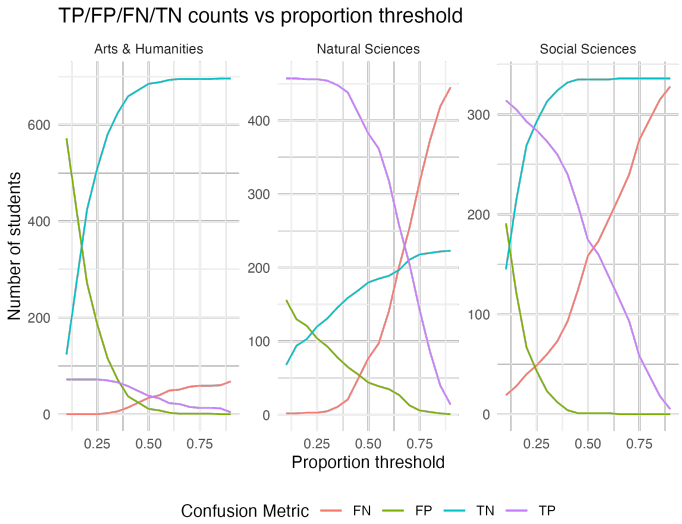

Figure 7: This plot details false negatives (FN), false positives (FP), true negatives (TN), and true positives (TP) for each of the divisions at different thresholds. In the Natural Sciences and Social Sciences there is a rapid rise in false negatives, indicating that too high of a threshold misses students who are majoring in these divisions. For Arts & Humanities, the number of students who truly are majoring in Arts & Humanities is low enough such that the penalty for false negatives is negligible.

Figure 8: Results from running imputation separately on each student group reveal only slight changes in the best threshold and resulting minimum error rates. There does not immediately appear to be trends across student groups.

Table 1: Imputed student major divisions for students matriculating in 2022 and 2023

predicted_divisions	n
Natural Sciences	1097
Social Sciences	1046
Arts & Humanities	195
Natural Sciences, Social Sciences	191
Arts & Humanities, Social Sciences	90
Arts & Humanities, Natural Sciences	17

The distribution of imputed majors is shown in Table 1. While my method of imputation successfully imputed nearly half of all missing major divisions; 9.7% of all observations in my modeling dataset still had missing values for student division. For these values, I resorted to the simple solution of filling with the mode for each of the three indicators. The result is a baseline of a Natural Science and Social Science major(s). Unfortunately, this adds systematic error to my dataset, although I deem the data I would have otherwise lost as of greater importance. At this point, no missing values remain.

5.2 Class Imbalance

An important detail to note about my data is that there is a large class imbalance. After removing courses for which 99% of students received no A-F grades, only 2% of rows were positive for S/U. This means that the overwhelming majority of 98.0% of rows were taken A-F (or the course was incomplete, withdrawn, taken as an audit, etc.). In this case, the best model to minimize error would chose to predict that courses are not taken on S/U basis. Based on a glimpse of the data, class imbalance marginally improved over time with more recent years having a slightly higher proportion of S/Us. Since my goal is primarily to evaluate differences before and after the pandemic, this evidence supported my decision to narrow my dataset to Fall 2015-Spring 2025. This improved imbalance to 97.6% vs 2.37%.

Class imbalance is not an unique problem. In fact, class imbalance problems have become more and more prevalent in literature with an increase in the number of published dissertations on imbalanced learning from 53 in 2013 to 521 in 2022 (Chen et al., 2024). The literature suggests data-level approaches such as oversampling of the minority class and undersampling the majority class as well as algorithm-level approaches of manipulating weights, adjusting decision thresholds, and using ensemble models (Chen et al., 2024) (Japkowicz & Stephen, 2002). Each technique has its own set of applications and drawbacks and various hybrid techniques also exist.

For this thesis, I explore the utilization of Synthetic Minority Over-sampling Technique (SMOTE) to address class imbalance. As the name suggests, SMOTE attempts to over

sample the minority class by generating synthetic data points based on existing samples and nearest neighbors (Chawla et al., 2002). Some common drawbacks of SMOTE include potential overfitting, performance overheard from the addition of observations, and mismatch between synthetic data and actual distribution of minority class. Notably, SMOTE is over 20 years old and in that time over 85 variants of SMOTE have been proposed (Fernandez et al., 2018). For simplicity, this thesis uses SMOTENC (extension for nominal and continuous features) proposed by Chawla et al. (2002) through `themis::step_smotenc` in the tidymodels ecosystem—although with limited success.

5.3 Hyperparameter Tuning

Before fitting models, I split my data into training (80%) and test (20%) sets. Data was split on student IDs to prevent data leakage; information about some of a student's courses likely reveals information that helps predict other observations by the same student. For modeling, I use the `tidymodels` ecosystem. While I intend to use random effects in my final model, preliminary model selection is performed without random effects due to computational limitations. Hence, I begin with logistic regression using `glm()` where the response variable is whether or not the course was taken S/U by a particular student.

The hyperparameters I experimented with tuning in this thesis are: over ratio, penalty, mixture, and decision threshold. Over ratio is used by SMOTE to determine how many synthetic samples should be generated. An over ratio of 1 tells SMOTE to generate samples until the balance between the minority and majority class is 1:1 while an over ratio of 0.5 tells SMOTE to generate samples until the ratio is 1:2. Penalty refers to regularization strength: how much to penalize model complexity. Mixture is used in elastic net regression to control the balance between LASSO and ridge regression. Generally, LASSO is used to minimize the number of predictors, whereas ridge corrects for overfitting and multi-collinearity. Finally, the decision threshold refers to the cutoff for classification. It answers the question: for which predicted probability values should S/U or not S/U be output.

All tuning of hyperparameters was done with 5-fold cross-validation. Over ratio (0.2 to 1), penalty (10^{-4} to 10^{1}), and mixture (0 to 1) were tuned together, testing each combination of the three hyperparameters. Five levels were tested for each hyperparameter, resulting in fitting the workflow 125 times. Afterward, I experimented with lowering the decision threshold to encourage more predictions of S/U. The model was tested with thresholds from 0.01 to 0.5 at 0.01 interval increments. The resulting best combination of hyperparameters was an over ratio of 0.4, penalty of 0.0001 (small penalty), mixture of 1 (pure LASSO), and decision threshold of 0.5.

5.4 Feature Selection

There are some variables provided that I did not include in my model. There was an indicator for whether or not a course was cross-listed that I dropped due to an inability to determine the nature of the crosslist (within-division, outside-division, outside Trinity). Instead of including the sparse data on specific majors, minors, and certificates, I engineered features (number of academic plans and division of major) as described previously in the Feature Engineering section. While I used class component to filter out courses, I opted to exclude it from modeling due to the lack of clarity I have on consistency across divisions and how credits are distributed for courses with a required lab. Additionally, what is a discussion section in one department could be equivalent to a lab in another. While there are additional specifications for a course to qualify as a seminar, I believe that a significant portion of that information comes from the number of students enrolled (cutoff at 18 students). I expect variation in class components to result in difficult to interpret results and leave this as an area of future study. Similarly, I did not explore the addition of interaction terms and transformations with one exception: the interaction between a student's major division and a course's division. This interaction term is of critical importance to my analysis as whether or not a student is taking a course as part of their major likely influences choice of grading—most majors do not permit courses taken S/U to fulfill major requirements.

At this point, I believe that all remaining variables in my modeling dataset have some relevance to student behavior. To confirm this belief, I turn towards feature selection strategies. Common techniques for variable selection include all-subset selection, step-wise AIC/BIC, LASSO, ridge regression, and elastic search. For this thesis, I did not explore all-subset selection or step-wise methods. All-subset selection would require high computational resources to compare each combination of features (2^p); step-wise methods are flawed due to their greedy approach being dependent on starting values and the lack of guarantee of an optimal solution. Hence, here I focus on elastic-net regression, a combination of LASSO and ridge regression. In addition to elastic–net, I isolated several predictors which I did not expect to have a strong correlation with voluntary S/U. I tested nested models with and without study away status (a large number of missing values were filled with FALSE), summer course status, number of academic plans (flaws in imputation for post-pandemic students), and my interaction terms. For the four colinear variables representing course load, I opted to compare AIC (metric of predictive accuracy) and BIC (how likely the model is to be the "true" model) for the full model with only one of the load variables.

From my run of elastic-net, all variables were kept. Similarly, the likelihood ratio tests I performed supported the inclusion of all variables in my final model with significant p-values ($<$ 0.05). AIC and BIC metrics (Table 2) also provide evidence in favor of keeping all variables. Notably, despite the deterministic relationship between load categories and load units, all four indicators of load were kept by elastic-net. This may suggest non-linearity in the relationship between load and the likelihood of opting for voluntary S/U. Alternatively, the inclusion of all load variables may simply be due to the size of my dataset having immense power to detect

small, uninteresting effects. For the sake of interpretability, I make the decision to select categorical indicators of load (load_stat, act_load) over their numeric counterparts (term_units, act_units). The categorical indicators are more representative than their numeric counterparts as course units are discrete and there is an imbalance when using the numeric values—a load of exactly 4.25 is rarer than one of 4.5.

Table 2: AIC and BIC metrics excluding select variables with SMOTENC, results are comparable without. Inclusion of all variables improves model fit.

Model	AIC	BIC	delta_AIC	delta_BIC
Full Model				
All_variables	142012	142340	0	0
Testing Questionable Variables				
No_took_summer_courses	142114	142432	102	92
No_num_students	142221	142540	209	199
No_studied_away	142266	142594	254	254
No_interaction	142367	142633	355	293
No_num_plans	142486	142794	474	454
Testing for Load Variable(s)				
Only_actload	142577	142864	565	524
Only_actunits	142642	142929	630	588
Only_termunits	142686	142973	674	633
Only_loadstat	142753	143051	741	710

5.5 Final Model

After addressing major modeling hurdles, I used the selected features and hyperparameters while adding a random effect on student ID using `lme4::glmer()`. The random effect serves to address violations of independence due to the repetition of students across the dataset. For instance, voluntary S/U courses are capped at 4.0 credits to count towards graduation requirements and students can only request voluntary for a single 1.0 credit course.

6 Results

6.1 Testing SMOTENC

When fitting the logistic regression without the random effect on student ID and using SMOTENC, the following diagnostics were generated. From Table 3, it is clear that SMOTENC did have an impact on the number of predictions of the minority class. When testing without

SMOTENC, no observations of the minority class were made. However, the quality of the predictions still leaves much to be desired with an abundance of false positives (predicting S/U but S/U was not chosen). This issue of overprediction is also visible in the calibration plot in Figure 9. For instance, when my model believes the probability of voluntary S/U is 0.1, the actual observed probability is only 0.013.

Table 3: Confusion matrix from running SMOTENC without random effect

	Actual	
	0	1
Pred.		
0	42896	910
1	1180	149

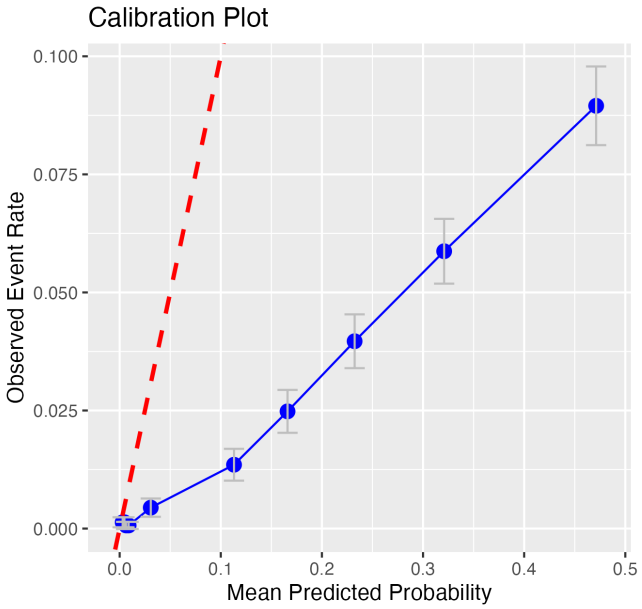

Figure 9: Calibration plot from running logistic regression with SMOTENC without random effects. The dashed red line represents a perfect model, the solid blue line represents my model, and the bars in gray are 95% confidence intervals.

After preliminary modeling without a random effect on student ID, mixed models were fit both with and without SMOTENC. However, SMOTENC did not appear to boost the minority class as neither model predicted any instances of S/U. This may be due to the inability for SMOTENC to understand the underlying hierarchical structure. SMOTE still treats the random effect as it does any other variable, synthesizing new student-course pairs. Unfortu-

nately, there appears to be a gap in the literature regarding the impact of using SMOTENC on generalized linear mixed models. Due to the challenges of interpreting SMOTENC in the context of random effects as well as the evident overconfidence when using SMOTENC on the general linear model, I present results without the use of class imbalance techniques. While hyperparameters were re-tuned and did vary, in the end the same decision threshold and features were chosen.

6.2 Logistic Regression with Student Effect (No SMOTENC)

DHARMa is specifically designed to check model assumptions for generalized linear mixed models (natively supports `glmer`) by generating multiple simulations and scaling residuals so they can be interpreted easily. As opposed to binned residuals, DHARMa is aware of the underlying hierarchical structure of the random effects. Figure 10 shows the diagnostic plots generated using DHARMa and Table 4 shows statistics and significance. On the left of Figure 10, the QQ plot checks for uniformity of scaled residuals and corresponds to the Kolmogorov-Smirnov (KS) uniformity test (Table 4). A KS statistic of 0 represents perfect uniformity while (heuristically) values above 0.1 likely indicate model misfit. While the p-value is significant, it is important to note that it is extremely sensitive to large sample sizes. Hence, from the low KS statistic of 0.004 and the near ideal pattern in the QQ plot, I do not believe there were violations in uniformity of scaled residuals—a signal that errors are random and evenly distributed. On the right of Figure 10, the scaled residuals versus fitted values appear to be evenly distributed above and below the y-axis. Residuals appear heavily clustered at towards 0, with fewer observations above 0.10. This is a feature of the heavily imbalanced dataset and does not cause me great concern as my intent is to identify correlates with voluntary S/U rather than suggesting causation or predicting S/U. Using the `performance` package, Table 4 also includes an estimate of intraclass correlation (ICC). Typically, adjusted ICC measures the proportion of variation in the data explained by the random effect. Assuming that ICC can be interpreted in the same way for GLMMs, 31.8% of observed variation can be explained by student ID.

If my goal was to predict voluntary S/Us, my model performs poorly. Results in Table 5 show that my model never predicts the rare case of voluntary S/U. From the calibration plot in Figure 11, we see that the model systematically underestimates the positive class (choosing voluntary S/U). However, when compared with the preliminary trial without random effects but using SMOTENC (Figure 9), it appears that the oversampling of the minority class has the side effect of over-prediction of S/Us—likely due to the misrepresentation of the true frequency of the event as an effect of oversampling. With the caveat that these are fundamentally two different models (one with random effects and one without), if I was to compare absolute deviance from the observed event rates, the degree of over-prediction from oversampling using SMOTENC seems to be more severe than the under-prediction without oversampling. Returning to the calibration plot in Figure 11, there does appear to be some non-linearity. For probabilities near zero, the model may be slightly over-predicting before

under-predicting (0.01 to 0.035). Towards 0.06, the model appears to begin returning back in the direction of over-prediction. This could be a result of non-linearity in some numeric predictors. For instance, perhaps the difference between having zero overloads and having one overload is different from the difference between three and four overloads.

As a supplement, I also plot ROC and PR curves (Figure 12) and compute the area under the curves (AUC) (Table 4). The ROC and its associated AUC support my model (PR AUC = 0.823) as it clearly performs better than a random model would (PR AUC = 0.5). While the PR curve and its associated AUC are very low (PR AUC = 0.083), the baseline for PR AUC is relative to the positive class's prevalence. A completely random model would have a baseline performance of 0.02. Thus, while a low bar, my model performs better than random—despite never predicting a positive. Due to the lack of positive predictions, I omit the typical analysis of precision, recall, specificity, and F1 score.

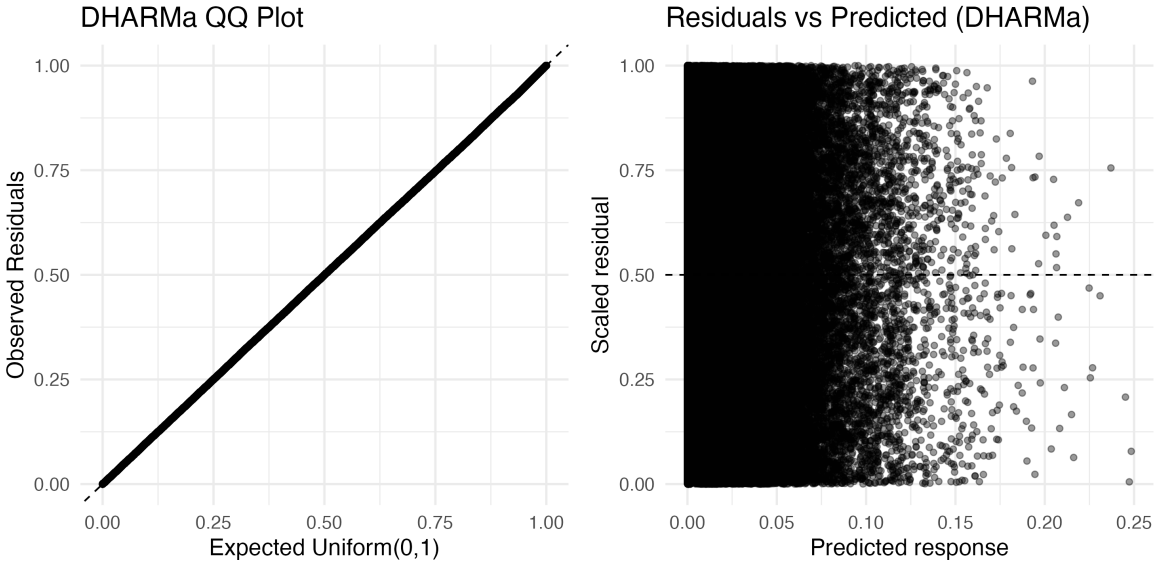

Figure 10: Model diagnostic plots where dotted lines represent ideal results.

Table 4: Model statistics from model with random effect (DHARMa, performance, and yardstick packages)

metric	estimate	p_value
KS test for uniformity	0.004	0.013
Adjusted ICC	0.318	–
ROC AUC	0.823	–
PR AUC	0.083	–

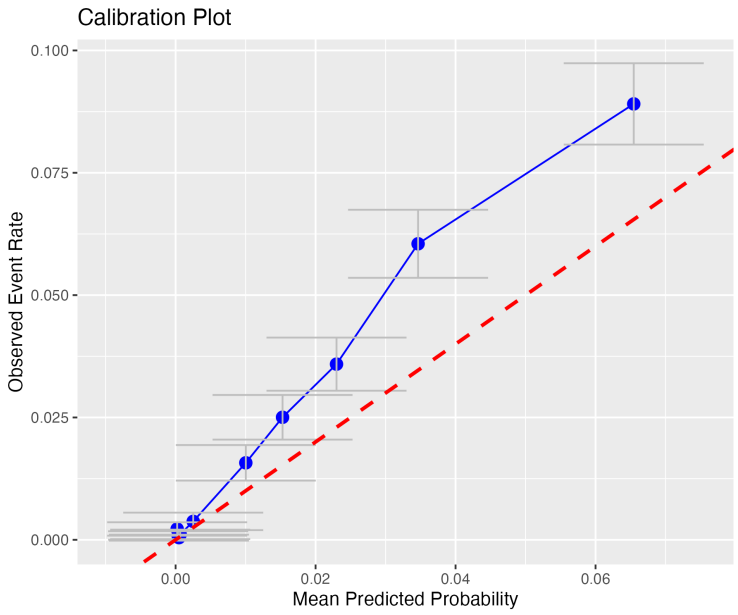

Figure 11: Calibration plot from running logistic regression with random effects on students (without SMOTENC). The dashed red line represents a perfect model, the solid blue line represents my model, and the bars in gray are 95% confidence intervals.

Table 5: Confusion matrix from running with random effect (no SMOTENC)

		Actual	
		0	1
Pred.			
	0	44076	1059
	1	0	0

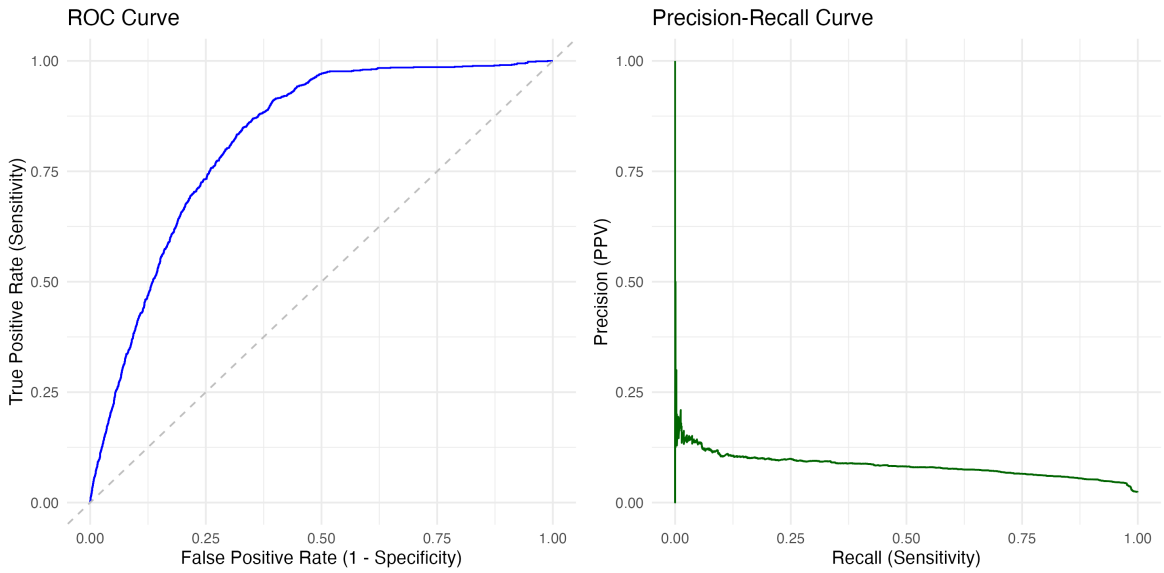

Figure 12: ROC curve and precision-recall curves for the full generalized linear mixed model. The dotted gray line in ROC graph represents a random classifier.

7 Discussion

From my model estimates (Table 6), I find several covariates to be statistically significant (p-value < 0.05). Levels of variables are ordinal and can be interpreted in order. For catalog level, note that they range from 0-99 to 400-499. The baseline represents a senior with a single major taking a 400-499 level course in Arts & Humanities before the pandemic. While unrealistic, they have not taken summer courses and have zeros for numeric variables (i.e. previous semester GPA, number of overloads).

7.0.1 Time Period Relative to COVID-19

My primary interest is in how COVID-19 induced policy changes impacted student choice of voluntary S/U. From my model estimates in Table 6, I find that all levels of time period are statistically significant. Notably, the magnitude of change in odds ratio (OR) for the time periods are the only ones my model finds to be greater than a factor of two. The baseline represents the pre-pandemic period (Fall 2015-Fall 2019). During the pandemic (Spring 2020-Spring 2022), there was approximately a 7-fold increase in odds of voluntary S/U relative to the baseline (OR [95% confidence interval] = 7.253 [6.530, 8.057]). For the post-pandemic period (Fall 2022-Spring 2025), there is an associated approximately 5-fold increase in odds of voluntary S/U relative to the baseline (OR = 5.196 [4.630, 5.832]). These results match my expectations from my exploratory analysis (Figures 1 and 2) which seemed to indicate a low

proportion of S/Us pre-pandemic, a drastic increase in S/U grades during the pandemic, and a slight decline in S/U grades after the pandemic—although still higher than pre-pandemic rates.

7.0.2 Summer Courses

Having taken summer courses is associated with a 1.065 [1.011, 1.121] increase in odds of voluntary S/U. Reasons for taking summer courses vary widely but may include retaking a course, not having alternative summer plans, and wanting to dedicate time to a challenging course. For any of these cases, the ultimate outcome is that, assuming passing grades, the student has fewer credits left to fulfill for graduation requirements. By default, without transfer credits, Trinity students must overload in order to meet the 34.0 credit graduation requirement (4.0 per semester for 8 semesters = 32.0 credits). One hypothesis for this effect is that having received some credits over the summer and hence may open the doors for these students to take fewer credits during spring and fall terms. If this is the case, students may be less incentivized to opt for S/U as their course load is more manageable than if they were to have been enrolled in an overload. Alternatively, perhaps taking summer courses is indicative of a different factor not captured in my dataset such as financial aid or pre-med status. However, my methodology is insufficient to determine causality and more work will have to be done to make a conclusion as to causal relationships; these are merely ideas.

7.0.3 Course Load

The relationship between overloading and voluntary S/U is partially captured in my model. For each additional semester of overloads a student enrolls in, their odds of taking a voluntary S/U in the subsequent semesters decreased by a factor of 0.828 [0.780, 0.879]. Similar to my hypothesis regarding summer sessions, one reasoning may be that having already overloaded may mean that in subsequent semesters there is less need to overload and hence less stress—which I hypothesize is a major reason to opt for voluntary S/U. More overloads may mean that a given student has already used up their four voluntary S/Us which may count towards graduation requirements. For students who have exceeded the number of overloads in order to meet credit requirements, perhaps this indicates a specific group of students who like to challenge themselves academically and may thus view voluntary S/U as a policy that contradicts their goals.

When taking into consideration a student's current course load, I observe that overloading is associated with a 0.857 [0.813, 0.905] factor decrease in odds of voluntary S/U. While estimates for academic load (calculated by excluding courses such as activity courses) were not statistically significant and hence not appropriate to extrapolate from, it is interesting to observe an opposite estimate. When students were enrolled in an academic overload, on average their odds of voluntary S/U slightly increased; when students were enrolled in an academic underload, on average their odds of voluntary S/U slightly decreased.

7.0.4 Student Level and Course Level

When compared to individuals with the academic level of a senior, all other academic levels had lower odds of voluntary S/U. First years, sophomores, and juniors had an odds ratio of 0.669 [0.621, 0.721], 0.769 [0.727, 0.812], and 0.797 [0.751, 0.845] respective to seniors. The strength of this effect may be overestimated as my data cleaning would have filtered out any seniors who graduated early as well as excluded semesters where students had part-time status—which is most prevalent among seniors in their last semester. Of the seniors who remain, the effect of their peers being part-time and the phenomenon of "senioritis" may hence result in an increase of voluntary S/Us. Furthermore, in their final semesters, students likely have a better understanding of how many S/U credits they have left for graduation and can take advantage of them without fear they may need them in the future. Alternatively, this could be indicative of a pattern wherein students tend to finish courses they cannot take voluntary S/U (for their major, minor, or certificate requirements) prior to senior year.

There is also a known and expected relationship between academic level and course levels as students are expected to begin with introductory courses and then advance onto higher level courses. Comparing lower catalog levels to 400-499 level courses, there is evidence of an increase in odds of voluntary S/U in all other catalog levels. Interestingly, the change in odds is not monotonic. The increase in odds when taking a 0-99 level course is a low 1.115 [1.058, 1.176], likely due to this category being mostly comprised of courses open only to first-year students. For 100-499 level courses, there is a trend of decreasing odds with increasing course level. Relative to 400-499 level courses, 100-199 level courses have the greatest increase in odds at 1.698 [1.547, 1.865], followed by 200-299 with 1.303 [1.198, 1.417] and 300-399 with 1.550 [1.409, 1.706]. This inverse relationship between course level and odds of voluntary S/U, excluding 0-99, signals that most voluntary S/Us occur at an introductory course level. This corresponds with the intent of voluntary S/U to be used for exploration of new topics. In part, this may also be a symptom of most degrees and departments prohibiting or restricting the use of S/U for degree requirements or in high-level courses.

7.0.5 Student Division and Course Division

In terms of course division, the Arts & Humanities have the lowest odds of being taken voluntary S/U. Relative to the Arts & Humanities, Natural Science and Social Science courses have relative odds of 1.374 [1.198, 1.577] and 1.169 [1.005, 1.359], respectively. This discrepancy between divisions may be from differences in expectations, grading scales, professors, culture, or other factors across divisions. In terms of post-undergraduate career, pre-med students generally must take their pre-med courses (mostly Natural Science but also some Social Science courses) on an A-F graded basis for their medical school applications. In contrast, it is perhaps more likely that for students in the Arts & Humanities or Social Sciences that their future is more comprised of going directly into the industry. This hypothesis is also supported by the estimates based on student major(s). Having a major in the Natural Sciences—most

common choice for pre-med students—greatly decreased the odds of voluntary S/U (OR = 0.722 [0.648, 0.805]). There was no significant relationship between majoring in Social Science or Arts & Humanities and the odds of voluntary S/U.

When looking at the interaction between student and course divisions, I expected to see same student division and course division pairs to have lower odds of voluntary S/U due to restrictions on voluntary S/U for major requirements. However, my model only statistically supports this hypothesis for Social Science. For Social Science majors, the odds of taking a Social Science course voluntary S/U decreased by a factor of 0.870 [0.774, 0.976]. While estimates for Natural Science students taking a Natural Science and Arts & Humanities students taking an Arts & Humanities course indicate a drop in voluntary S/U, they are not statistically significant. Perhaps this discrepancy between divisions may be indicative of differential underlying within-division differences. For instance, Arts & Humanities includes a wide range of departments from Brazil and Global Portuguese Studies to Philosophy to Dance. Although Natural Sciences is composed of a smaller number of departments, a wide range of careers may stem from the range from Mathematics to Global Health to Evolutionary Anthropology. In a similar vein, for different student major division and course division pairs, I expected to see higher odds of voluntary S/U as these courses may be taken on a more exploratory basis with only a few courses that may count towards major requirements. However, for most cross-divisional pairs, there is no statistically significant difference in odds of voluntary S/U. Only for Natural Science majors taking a course in the Arts & Humanities was there a statistically significant 1.160 [1.061, 1.269] factor increase in odds of voluntary S/U. This may simply be a result stemming from the rarity of the combination of Natural Science and Arts & Humanities (Table 1). In which case, to fulfill general education requirements, only a very small number of students will be using their Arts & Humanities credits for another academic plan, incentivizing voluntary S/U.

7.0.6 Other Significant Covariates

While holding all other features constant, for every one unit (1.0) increase in previous term GPA, the odds of a student electing to take a voluntary S/U decreased by a factor of 0.942 [0.904, 0.980]. This relationship may indicate that having a better GPA in the previous semester decreases the incentive to resort to more lenient S/U grading in the following semester.

In terms of the number of academic plans, my model suggests that the difference between having a single major and having a single major in addition to another plan (major, minor, or certificate) is on average associated with an increase in odds of voluntary S/U by a factor of 1.166 [1.091, 1.246]. However, this should be taken with caution as there is likely a disproportionately low number of plans in my dataset as some students have not graduated and may choose to add plans later in their Duke experience. The non-significant and negligible difference in having three academic plans may similarly be a result of a lack of observations.

Table 6: Fixed effects from final model

term	Odds Ratio (OR)	OR_low	OR_high	P-Value
(Intercept)	0.004	0.003	0.004	0.000
Time Period				
timeperiod_dur_covid	7.253	6.530	8.057	0.000
timeperiod_post_covid	5.196	4.630	5.832	0.000
Course Load				
load_status_underload	0.967	0.932	1.003	0.075
load_status_overload	1.047	0.987	1.111	0.128
actual_load_overload	0.857	0.813	0.905	0.000
num_overloads	0.828	0.780	0.879	0.000
Student Level				
academic_level_first_year	0.669	0.621	0.721	0.000
academic_level_sophomore	0.797	0.751	0.845	0.000
academic_level_junior	0.769	0.727	0.812	0.000
Course Level				
catalog_level_0-99	1.115	1.058	1.176	0.000
catalog_level_100s	1.698	1.547	1.865	0.000
catalog_level_200s	1.550	1.409	1.706	0.000
catalog_level_300s	1.303	1.198	1.417	0.000
Student Division				
is_art_humanity	1.009	0.917	1.110	0.859
is_social_sci	0.980	0.903	1.064	0.629
is_natural_sci	0.722	0.648	0.805	0.000
Course Division				
division_Natural.Sciences	1.374	1.198	1.577	0.000
division_Social.Sciences	1.169	1.005	1.359	0.042
Division Interactions				
is_natural_sci_x_division_Arts...Humanities	1.160	1.061	1.269	0.001
division_Social.Sciences_x_is_social_sci	0.870	0.774	0.976	0.018
is_natural_sci_x_division_Natural.Sciences	0.921	0.820	1.034	0.165
division_Arts...Humanities_x_is_art_humanity	0.973	0.897	1.056	0.515
division_Arts...Humanities_x_is_social_sci	1.047	0.966	1.134	0.265
division_Natural.Sciences_x_is_art_humanity	1.008	0.951	1.068	0.799
Miscellaneous				
prev_semGPA	0.942	0.904	0.980	0.004
took_summer_courses	1.065	1.011	1.121	0.017
num_plans_2	1.166	1.091	1.246	0.000
num_plans_3	1.000	0.931	1.075	0.994
num_students	1.011	0.971	1.053	0.604

7.1 Limitations

As with most studies, there are numerous limitations with this thesis. To begin, there is systematic missingness of academic plans for Fall 2022-Spring 2025. There are certainly misclassifications in my imputation of student major divisions and for students still early in their academic career and unable to be imputed, the most common value was assigned for each of the three major division indicators (Social Science and Natural Science double major). However, my model does not know the underlying restriction that a student cannot have all three major divisions—you cannot do an interdepartmental major along with a second major. In many cases, students realize later in their academic careers that they have the time to explore other interests. As a result, I make no attempt to impute students' minors or certificates and their number of plans are likely an underestimate. Since these students have not had four academic years, it is also impossible to know if they will drop out or graduate early/late.

Missing data is not exclusive to after the pandemic. There was also some missing data on course location and the impossibility of having data on previous semesters for first semester students. The choice to systematically give all students a 4.0 as the baseline previous semester GPA for their first semester decreases the perceived significance of prior GPA in my model. There is no simple solution to this as there is a lack of knowledge of a student's prior academic performance in earlier education. This is exacerbated by the loss of the standardized test score application requirement for Duke University and would result in only the usage of high school GPAs, which will likely vary in rigor across schools. It is also worth emphasizing that while my engineered previous semester GPA does consider courses taken away from Durham, it ignores summer terms which do contribute to a student's overall GPA. My major imputation approach currently ignores both summer and study away courses.

Beyond the variables I was provided, it is likely that ROTC, pre-med, athlete, and other statuses may have an impact on student behaviors. For instance, it makes sense to believe that ROTC and athlete students may be taking a disproportionate number of "Other" credits to fulfill program requirements. Students intending to apply for medical school have to take a significant number of courses in the natural sciences, but may not necessarily be majoring in the natural science division. Related to these unknowns is my systematic ignorance of student minors and certificates. It is certainly possible that minors in a separate division will fully account for all remaining graduation requirements. Other variables not captured in my dataset include other policies and policy changes that were not explicitly denoted in the Duke Bulletins or otherwise made aware to me. It is possible that I neglected changes in which courses are allowed to be taken voluntary S/U and by whom, or I may be misinterpreting the language used by the Trinity College. I am aware that I did not verify that all student-course pairs used in my model satisfy the requirement of four 1.0 credit courses, as currently stated for the 2025-2026 academic year (*Courses*, n.d.).

In the dataset I was given, there are also some potentially inconsistent observations and unclear methods. As mentioned in the Data section, there may be some errors in the recorded course enrollments as well as discrepancies between number of semesters and internal academic level

calculations. The algorithm for computing a student's academic level also remains unknown to me. Beyond that, it is also plausible that a course is cross-listed across divisions. For instance, a course could be taught on the interaction of music (arts & humanities) and neuroscience (natural science), counting as an elective for both departments. Yet, in my dataset only one of these divisions is displayed.

Some methods used in this thesis are chosen for the sake of simplicity rather than having a clear motive. For instance, the labeling of "traditional" students and "academic" courses are based on personal beliefs. Perhaps I have unconscious biases against Arts & Humanities by removing half-credit music courses. However, I also believe there are systematic biases in the classification of departments into divisions. As mentioned in my methodology, the classification of African & African American Studies as a Social Science and the classification of Asian & Middle Eastern Studies as an Arts & Humanities department seems to neglect the interdisciplinary nature of many fields. In terms of modeling approach, I make assumptions in order to fit the logistic regression. Notably, I assume that predictors are linear with the log odds of the outcome (choice of voluntary S/U). Yet, based on ridge regression's inclusion of both course units and the categorical load, there may be some areas of nonlinearity.

Finally, since my model seeks to determine student-level behavior rather than course-level behavior, I do not include course IDs in my model. Hence, I do not take into consideration the potential for some courses to be known within the student population as a popular course for graduation requirements. For instance, ECS 101 is known as "Rocks for Jocks" among the student community and could therefore have a disproportionate proportion of S/Us when compared to ENVIRON 101, a course which would likely have identical values in my dataset (excluding total enrollment). Similarly, it is plausible that a specific instructor and/or a specific course has a different stance towards requirements for S/U and grade cutoffs than their peers.

8 Conclusion

The COVID-19 pandemic precipitated profound disruptions in higher education, prompting rapid and, in many cases, unprecedented changes to academic policies. Among these was the expansion of flexible grading options, including the temporary broadening of Satisfactory/Unsatisfactory (S/U) grading at Duke University. Although originally implemented as an emergency response, the durability of certain components of these policies has raised questions regarding their long-term implications for student behavior. This thesis provides empirical evidence that the pandemic-era shift in grading policy has had sustained effects on students' voluntary use of S/U grading.

Drawing on administrative records from the Duke University Assessment Office, I employed a logistic mixed-effects modeling framework to examine student decisions regarding voluntary S/U grading across nearly a decade of academic terms. After accounting for course characteristics, student academic history, and individual-level heterogeneity, the analysis reveals

a statistically meaningful increase in voluntary S/U selection in the post-pandemic period relative to pre-pandemic norms. The findings suggest that exposure to the exceptional grading environment of Spring 2020 and the subsequent pandemic-affected semesters may have altered patterns of student decision-making in ways that persist even under largely restored instructional conditions.

The significance of predictors such major division, course division, overload status, course level, student academic level, previous term GPA, and having taken summer courses indicate the multifaceted nature of voluntary S/U usage. Even so, the persistent effect of the post-pandemic indicator suggests that structural conditions alone cannot fully account for the observed behavioral shift. These results carry implications for ongoing institutional deliberations regarding the future of flexible grading. If voluntary S/U usage is now embedded in student behavior rather than merely a temporary artifact of crisis conditions, then policy frameworks created for pre-pandemic educational environments may be insufficiently aligned with current practices. As departments reconsider whether and how S/U grading should be permitted, the evidence presented here may assist in evaluating the trade-offs among flexibility, academic rigor, signaling value, and student well-being.

8.1 Future Work

This study also highlights the need for further research. For the improvement of the work carried out in this thesis, a couple of items come into mind. To begin, some directly implementable actions relevant to this thesis could include the tuning of additional hyperparameters such as the number of neighbors used for SMOTE, further feature engineering (overall cumulative GPA instead of previous semester GPA), exploration of variable transformations and interaction terms, and experimentation with other packages such as `glmmTMB`. It may also be worth modeling on a smaller, more specific dataset that could have less imbalance. There are also many alternative approaches to missing data that could be experimented with. For instance, my current major imputation approach only considers the proportion of courses a student takes in each division above 199. An alternative could be to use a package like `mice` (Multivariate Imputation by Chained Equations) which could capture other informative covariates for major (van Buuren & Groothuis-Oudshoorn, 2011). Additionally, my methodology lacks the enforcement of constraints such as the impossibility of a student majoring in all three divisions. More complex adjustments include exploration of an even more complex yet more precise hierarchical structure (perhaps with a Bayesian approach) that allows variation in student behavior at the division level, the department level, and at the individual level.

Most distal extensions to this thesis may consist of the use of different methods to address class imbalance and imputation. For class imbalance, techniques may include: weights, alternative variations of SMOTE, models resistant to imbalance (e.g. Support Vector Machine) (Prati et al., 2014), propensity score matching (Austin, 2011), and other approaches—and their hybrids. Developments in methods for class imbalance will likely also have extensions in a variety of other fields as well. For major imputation (or prediction), the problem has been

explored in more depth by others such as (Lang et al., 2022). It could be of interest to identify patterns in major selection at Duke University and compare them with wider shifts across the nation. A related problem is the inconsistent and somewhat arbitrary categorization of majors into divisions. It is impossible to categorize all of human knowledge into distinct, disjoint departments. One method for doing so could be to use natural language processing techniques on course descriptions and syllabi. By using vector embeddings and computing cosine similarities, one could approximate the relationships between departments and institutes (*Cosine Similarity*, n.d.).

8.2 Recommendations

One of the major obstacles in this thesis was the lack of post-pandemic data. I believe that future studies using actual student majors and using longer-term data may be able to provide a more accurate framework for evaluating factors influencing student behavior towards voluntary S/U. Moving even further forward, a retrospective analysis of how the shift to opt-in S/U grading on the instructor side and its utilization may also be worth noting. Similarly, the shift to the newest Duke undergraduate curriculum beginning in Fall 2025 should also be evaluated for its intersection with student behaviors.

While factors such as student demographics, extracurricular, scholarships, and health conditions are likely to remain inaccessible due to privacy and safety concerns, less sensitive features to explore may include course attributes associated with individual classes and the departments in which courses belong. At the root, the reason for my interest in specific academic programs (majors, minors, certificates) was that some programs lack courses that satisfy certain requirements under Curriculum 2020. For instance, I do not believe there has been a recent course listed in the statistical science department that has had the code for Arts, Literature, and Performance (ALP). Yet, I believe it would be quite feasible to have an applied statistics course on something such as learning natural language processing techniques and applying them to literature. A summary of the proportion of courses and their associated curriculum codes listed in each department could help identify interdisciplinary gaps to be filled. Additionally, there has been speculation that departments with higher average GPAs tend to attract more students than they would otherwise (Sabot & Wakeman-Linn, 1991). A summary of department level GPAs could help calibrate the difficulty of coursework to be more balanced across divisions and course levels.

Beyond flexible grading policies, I would also recommend that research be conducted on the use of audits. From historic Duke Bulletins, the Trinity College of Arts and Sciences began permitting academic audits in 1960 (Registrar, n.d.). It was only in 1966 that flexible grading was introduced with the pass/fail system. Audits serve a similar purpose of allowing academic exploration with even lower risk than flexible grading. However, they do not count towards course credit and hence do not necessarily decrease a student's overall workload—nor do they intuitively seem to be a cause of grade inflation. For administrators and educators, it may

be worth analyzing patterns in audit usage in order to weigh the trade-offs and use-cases for flexible grading versus audits.

Code Availability

Code used for this thesis can be found at https://github.com/sophiazyang/senior-thesis.

References

2019 NAEP high school transcript study (HSTS) results. (n.d.). National Assessment of Educational Progress. Retrieved November 11, 2025, from https://www.nationsreportcard.gov/hstsreport/#home

Academic Honors and Recognition. (n.d.). Office of the University Registrar, Duke University. https://registrar.duke.edu/student-resources/academic-honors-and-recognition/

Agostino, H., Burstein, B., Moubayed, D., Taddeo, D., Grady, R., Vyver, E., Dimitropoulos, G., Dominic, A., & Coelho, J. S. (2021). Trends in the Incidence of New-Onset Anorexia Nervosa and Atypical Anorexia Nervosa Among Youth During the COVID-19 Pandemic in Canada. *JAMA Network Open*, *4*(12). https://doi.org/10.1001/jamanetworkopen.2021.37395

American Academy of Arts & Sciences. (2014). *Public research universities: Changes in state funding*. https://www.amacad.org/sites/default/files/academy/multimedia/pdfs/publications/researchpapersmonographs/PublicResearchUniv_ChangesInStateFunding.pdf

Ange, B., Wood, E. A., Thomas, A., & Wallach, P. M. (2018). Differences in Medical Students' Academic Performance between a Pass/Fail and Tiered Grading System. *Southern Medical Journal*, *111*(11), 683–687. https://doi.org/10.14423/smj.0000000000000884

Ashbaugh, E. J., & Chapman, H. B. (1925). Report cards in american cities. *Educational Research Bulletin*, *4*(14), 289–293.

Austin, P. C. (2011). An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivariate Behav Res.*, *46*(3), 399–424. https://doi.org/10.1080/00273171.2011.568786

Banks, R. R., Levine, E. J., Olick Llano, E., Pham, H., Stevens, M. L., & Sutton, D. (2024). *Private Universities in the Public Interest – White Paper*. Stanford Center for Racial Justice at Stanford Law School. https://law.stanford.edu/stanford-center-for-racial-justice/projects/private-universities-in-the-public-interest/private-universities-in-the-public-interest-white-paper/

Beaver, W. (2017). The Rise and Fall of For-Profit Higher Education. *Academe*, *103*(1). https://www.aaup.org/academe/issues/103-0/rise-and-fall-profit-higher-education

Bejar, I. I., & Blew, E. O. (1981). Grade Inflation and the Validity of the Scholastic Aptitude Test. *American Educational Research Journal*, *18*(2), 143–156. https://doi.org/10.3102/00028312018002143

Beland, L.-P., & Kim, D. (2016). The Effect of High School Shootings on Schools and Student Performance. *Educational Evaluation and Policy Analysis*, *38*(1). https://doi.org/10.3102/0162373715590683

Best college rankings. (n.d.). U.S. News. https://www.usnews.com/rankings

Bettinger, E. P., & Long, B. T. (2010). Does Cheaper Mean Better? The Impact of Using Adjunct Instructors on Student Outcomes. *The Review of Economics and Statistics*, *92*(3), 598–613. https://ideas.repec.org/a/tpr/restat/v92y2010i3p598-613.html

Birnbaum, R. (1977). Factors Related to University Grade Inflation. *The Journal of Higher Education*, *48*(5), 519–539. https://doi.org/10.1080/00221546.1977.11776572

Bixler, H. H. (1936). School marks. *Review of Educational Research*, *6*(2), 169–173.

Bloodgood, R. A., Short, J. G., Jackson, J. M., & Martindale, J. R. (2009). A Change to Pass/Fail Grading in the First Two Years at One Medical School Results in Improved Psychological Well-Being. *Academic Medicine*, *84*(5), 655–662. https://doi.org/10.1097/ACM.0b013e31819f6d78

Blum, D. (2017). Nine Potential Solutions to Abate Grade Inflation at Regionally Accredited Online U.S. Universities: An intrinsic case study. *The Qualitative Report*, *22*(9). https://files.core.ac.uk/download/pdf/132324597.pdf

Brück, T., Di Maio, M., & Miaari, S. H. (2019). Learning The Hard Way: The Effect of Violent Conflict on Student Academic Achievement. *Journal of the European Economic Association*, *17*(5). https://doi.org/10.1093/jeea/jvy051

Burk, D., & Perry, J. (2020). *The Volume and Repayment of Federal Student Loans: 1995 to 2017*. Congressional Budget Office. https://eric.ed.gov/?id=ED610721

Centra, J. A. (1993). *Reflective Faculty Evaluation: Enhancing Teaching and Determining Faculty Effectiveness. The Jossey-Bass Higher and Adult Education Series*. Jossey-Bass Inc. https://doi.org/10.1080/00213624.2008.11507197

Chan, C. K. Y. (2023). A review of the changes in higher education assessment and grading policy during covid-19. *Assessment & Evaluation in Higher Education*, *48*(6), 874–887. https://doi.org/10.1080/02602938.2022.2140780

Chan, W., Hao, L., & Suen, W. (2007). A SIGNALING THEORY OF GRADE INFLATION*. *International Economic Review*, *48*(3), 1065–1090. https://doi.org/10.1111/j.1468-2354.2007.00454.x

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal Of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.48550/arXiv.1106.1813

Chen, W., Yang, K., Yu, Z., Shi, Y., & Chen, P. C. L. (2024). A survey on imbalanced learning: Latest research, applications and future directions. *Artificial Intelligence Review*, *57*(137). https://doi.org/10.1007/s10462-024-10759-6

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., & York, R. L. (1966). *EQUALITY OF EDUCATIONAL OPPORTUNITY*. U.S. Office of Education. https://eric.ed.gov/?id=ED012275

Collins, J. R., & Nickel, K. N. (1975). Grading policies in higher education: The kansas study/the national survey. In *University Studies* (Vol. 103). Wichita State University.

Collins, R. (1979). *The Credential society : An historical sociology of education and stratification*. New York : Academic Press. http://archive.org/details/credentialsociet0000coll

Cosine Similarity. (n.d.). ScienceDirect. Retrieved November 11, 2025, from https://www.sciencedirect.com/topics/computer-science/cosine-similarity

Courses: Satisfactory / Unsatisfactory Grading Option. (n.d.). Duke Trinity College of Arts & Sciences. Retrieved November 11, 2025, from https://trinity.duke.edu/undergraduate/academic-policies/unsatisfactory-satisfactory-grading-option

Cross, K., Evans, J., MacLeavy, J., & Manley, D. (2022). Analysing the socio-economic impacts of COVID-19: A new regional geography or pandemic enhanced inequalities? *Regional Studies, Regional Science*, *9*(1). https://doi.org/10.1080/21681376.2022.2084447

Crumbley, D. L., Flinn, R. E., & Reichelt, K. J. (2010). What is Ethical About Grade Inflation and Coursework Deflation? | Journal of Academic Ethics. In *Journal of Academic Ethics* (Vol. 8, pp. 187–197). https://link.springer.com/article/10.1007/s10805-010-9117-9

Curriculum for Students Enrolling in Fall 2025. (n.d.). Duke Trinity College of Arts & Sciences. Retrieved November 11, 2025, from https://trinity.duke.edu/undergraduate/academic-policies/curriculum-students-enrolling-fall-2025

Curriculum: Overview | Trinity College of Arts & Sciences. (n.d.). Duke Trinity College of Arts & Sciences. Retrieved November 11, 2025, from https://trinity.duke.edu/undergraduate/academic-policies/curriculum

Davidson, J. F. (1975). Academic Interest Rates and Grade Inflation. *Educational Record*, *56*(2), 122–125.

Dederichs, M., Weber, J., Muth, T., Angerer, P., & Loerbroks, A. (2020). Students' perspectives on interventions to reduce stress in medical school: A qualitative study. *PLOS ONE*, *15*(10). https://doi.org/10.1371/journal.pone.0240587

Devaney, T. A., Carr, S. C., & Allen, D. D. (2009). Impact of Hurricane Katrina on the Educational System in Southeast Louisiana: One Year Follow-Up. *Research in the Schools*, *16*(1), 32–44. https://www.researchgate.net/profile/Thomas-Devaney/publication/295702490_Impact_of_Hurricane_Katrina_on_the_Educational_System_in_Southeast_Louisiana_One_Year_Follow-Up/links/56cc9d5208ae059e37506abd/Impact-of-Hurricane-Katrina-on-the-Educational-System-in-Southeast-Louisiana-One-Year-Follow-Up.pdf

Diorio, G. L. (2023). History of Public Education in the U.S | Research Starters | EBSCO Research. In *EBSCO*. https://www.ebsco.com/research-starters/history/history-public-education-us

Duke coronavirus response. (n.d.). Duke University. Retrieved November 11, 2025, from https://coronavirus.duke.edu/updates/

Duke University: A Brief Narrative History. (2020). Duke University Libraries. https://library.duke.edu/rubenstein/uarchives/history/articles/narrative-history

Duke Welcomes The Newest Members of the Class of 2029. (2025). Duke Today. https://today.duke.edu/2025/03/duke-welcomes-newest-members-class-2029

Dyrbye, L. N., Thomas, M. R., & Shanafelt, T. D. (2005). Medical Student Distress: Causes, Consequences, and Proposed Solutions. *Mayo Clinic Proceedings*, *80*(12), 1613–1622. https://doi.org/10.4065/80.12.1613

Edmonds, R. R. (1979). Effective schools for the urban poor. *Educational Leadership*, *37*(1), 15–24. https://files.ascd.org/staticfiles/ascd/pdf/journals/ed_lead/el_197910_edmonds.pdf

Education, D. of. (2014). *Program Integrity: Gainful Employment*. https://www.ed.gov/sites/ed/files/policy/highered/reg/hearulemaking/2012/notice-proposed-rulemaking-march-14-2014.pdf

Ehlers, T., & Schwager, R. (2016). Honest Grading, Grade Inflation, and Reputation. *CESifo Economic Studies*, *62*(3), 506–521. https://doi.org/10.1111/j.1468-2354.2007.00454.x

Ehrenreich, B. (1989). *Fear of falling : The inner life of the middle class*. New York : Pantheon Books. http://archive.org/details/fearoffalling00barb

Ellis, L., Burke, D. M., Lomire, P., & McCormack, D. R. (2003). Student Grades and Average Ratings of Instructional Quality: The Need for Adjustment. *The Journal of Educational Research*, *97*(1), 35–40. https://doi.org/10.1080/00220670309596626

Elsner, P. A., & Brydon, C. W. (1974). *Nonpunitive Grading Practices and Policies*. https://eric.ed.gov/?id=ED088549

Fernandez, A., Garcia, S., Herrera, F., & Chawla, N. V. (2018). SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary. *Journal Of Artificial Intelligence Research*, *61*, 863–905. https://doi.org/doi.org/10.1613/jair.1.11192

Finkelstein, I. E. (1913). *The Marking System in Theory and Practice*. Warwick & York, Incorporated.

Gardner, D. P., & Others. (1983). *A Nation At Risk: The Imperative For Educational Reform. An Open Letter to the American People. A Report to the Nation and the Secretary of Education*. National Commission on Excellence in Education, U.S. Department of Education. https://eric.ed.gov/?id=ED226006

Gibbs, L. (2020). *#PassFailNation: Alternate Grading*. https://oudigitools.blogspot.com/2020/03/feedback-alternate-grading-in-crisis.html

Goldin, C. (2010). *Public school districts and elementary, secondary, and one-teacher schools, by public-private control: 1916–1996*. Cambridge University Press. https://doi.org/10.1017/ISBN-9780511132971.Bc1-509

Goldman, L. (1985). THE BETRAYAL OF THE GATEKEEPERS: GRADE INFLATION. *The Journal of General Education*, *37*(2), 97–121. https://www.jstor.org/stable/27797025

Hadley, S. T. (1954). A school mark-fact or fancy. *Educational Administration and Supervision*, *40*, 305–312.

Halper, D. (2025). Arts & Sciences Council approves changes to S/U grading, hears update on shifting collegiate athletics landscape. In *The Duke Chronicle*. https://dukechronicle.com/article/duke-university-arts-and-sciences-council-changes-satisfactory-unsatisfactory-grading-system-duke-athletics-changing-landscape-20250503

Hammerstein, S., König, C., Dreisörner, T., & Frey, A. (2021). Effects of COVID-19-Related School Closures on Student Achievement-A Systematic Review. *Frontiers in Psychology*, *12*. https://doi.org/10.3389/fpsyg.2021.746289

Hanson, M. (2025). *Educational Attainment Statistics [2025]*. https://educationdata.org/education-attainment-statistics

Haswell, R. H. (1999). Grades, Time, and the Curse of Course. *College Composition and Communication*, *51*(2), 284–295. https://doi.org/10.2307/359043

Hoyt, D. P. (1966). College grades and adult accomplishment: A review of research. *The Educational Record*, *47*, 70–75.

Iris Franz, W.-J. (2010). Grade inflation under the threat of students' nuisance: Theory and evidence. *Economics of Education Review*, *29*(3), 411–422. https://doi.org/10.1016/j.econedurev.2009.10.013

Japkowicz, N., & Stephen, S. (2002). The class imbalance problem: A systematic study. *Intelligent Data Analysis: An International Journal*, *6*(5). https://doi.org/10.3233/IDA-

2002-6504

Jessee, W. F., & Simon, H. J. (1971). Time utilization by medical students on a pass-fail evaluation system. *Journal of Medical Education*, *46*(4), 275–280. https://doi.org/10.1097/00001888-197104000-00003

Johnson, J. T. (1970). Evaluate program, not grading. *College and University Business*, *49*(3), 77–78.

Karlins, M., Kaplan, M., & Stuart, W. (1969). Academic Attitudes and Performance as a Function of Differential Grading Systems. *The Journal of Experimental Education*, *37*(3), 33–50. https://doi.org/10.1080/00220973.1969.11011129

Kuhfeld, M., Soland, J., & Lewis, K. (2022). Test Score Patterns Across Three COVID-19-Impacted School Years. *Educational Researcher*, *51*(7), 500–506. https://doi.org/10.3102/0013189X221109178

Kuperman, V., Geva, E., Taler, V., & Thériault, K. (2025). Recovery from university grade inflation after the COVID-19 pandemic varies by faculty. *Studies in Higher Education*, 1–16. https://doi.org/10.1080/03075079.2025.2470297

Lang, D., Wang, A., Dalal, N., Paepcke, A., & Stevens, M. L. (2022). Forecasting Undergraduate Majors: A Natural Language Approach. *AERA Open*, *8*(1), 1–18. https://doi.org/10.1177/23328584221126516

Langbein, L. (2008). Management by results: Student evaluation of faculty teaching and the mis-measurement of performance. *Economics of Education Review*, *27*(4), 417–428. https://doi.org/10.1016/j.econedurev.2006.12.003

Lezotte, L. W. (1991). *Correlates of effective schools: The first and second generation*. Effective Schools Products, Ltd. https://www.effectiveschools.com/Correlates.pdf

Love, D. A., & Kotchen, M. J. (2010). Grades, Course Evaluations, and Academic Incentives. *Eastern Economic Journal*, *36*(2), 151–163. https://ideas.repec.org//a/pal/easeco/v36y2010i2p151-163.html

Maldonado, C. (2018). *Price Of College Increasing Almost 8 Times Faster Than Wages*. Forbes. https://www.forbes.com/sites/camilomaldonado/2018/07/24/price-of-college-increasing-almost-8-times-faster-than-wages

Mitchell, M., Leachman, M., Masterson, K., & Waxman, S. (2018). *Unkept Promises: State Cuts to Higher Education Threaten Access and Equity | Center on Budget and Policy Priorities*. Center on Budget; Policy Priorities. https://www.cbpp.org/research/state-budget-and-tax/unkept-promises-state-cuts-to-higher-education-threaten-access-and

Mostafa, S. A., Ferguson, R., Tang, G., & Ashqer, M. (2023). An Analysis of the COVID-19-Induced Flexible Grading Policy at a Public University. *Higher Education Policy*, 1–34. https://doi.org/10.1057/s41307-023-00315-2

Mutch, C. (2014). The role of schools in disaster preparedness, response and recovery: What can we learn from the literature? *Pastoral Care in Education*, *32*(4). https://doi.org/10.1080/02643944.2014.880123

Neath, I. (1996). How to Improve Your Teaching Evaluations without Improving Your Teaching. *Psychological Reports*, *78*(3_suppl), 1363–1372. https://doi.org/10.2466/pr0.1996.78.3c.1363

Neville, R. D., Lakes, K. D., Hopkins, W. G., Tarantino, G., Draper, C. E., Beck, R., &

Madigan, S. (2022). Global Changes in Child and Adolescent Physical Activity During the COVID-19 Pandemic: A Systematic Review and Meta-analysis. In *JAMA Pediatrics* (9; Vol. 176). https://doi.org/10.1001/jamapediatrics.2022.2313

Noble, J. P., & Sawyer, R. L. (2004). Is High School GPA Better Than Admission Test - ProQuest. *College and University*, *76*(4), 17–22. https://www.proquest.com/docview/225613390?sourcetype=Scholarly%20Journals

O'Halloran, K. C., & Gordon, M. E. (2014). A synergistic approach to turning the tide of grade inflation. *Higher Education*, *68*, 1005–1023. https://doi.org/10.1007/s10734-014-9758-5

Perry, B. L., Aronson, B., & Pescosolido, B. A. (2021). Pandemic precarity: COVID-19 is exposing and exacerbating inequalities in the American heartland | PNAS. *Proc. Natl. Acad. Sci. U.S.A.*, *118*(8). https://doi.org/10.1073/pnas.202068511

Prati, R. C., Batista, G. E. A. P. A., & Silva, D. F. (2014). Class imbalance revisited: A new experimental setup to assess the performance of treatment methods. *Knowledge and Information Systems*, *45*, 247–270. https://doi.org/10.1007/s10115-014-0794-3

Prichett, L. M., Yolken, R. H., Severance, E. G., Carmichael, D., Zeng, Y., Lu, Y., Young, A. S., & Kumra, T. (2024). COVID-19 and Youth Mental Health Disparities: Intersectional Trends in Depression, Anxiety and Suicide Risk-Related Diagnoses. *Academic Pediatrics*, *24*(5), 837–847. https://doi.org/10.1016/j.acap.2024.01.021

Provan, J., & Cuttress, L. (1995). PREFERENCES OF - PROGRAM DIRECTORS FOR EVALUATION OF CANDIDATES FOR POSTGRADUATE TRAINING. *Canadian Medical Association Journal*, *153*(7). https://www.webofscience.com/wos/woscc/full-record/WOS:A1995RY10600020

Quann, C. J. (1971). *The pass/fail option: Analysis of an experiment in grading*. American Association of Collegiate Registrars; Admissions Officers; Paper presented at the 57th Annual Meeting of the American Association of Collegiate Registrars and Admissions Officers. https://files.eric.ed.gov/fulltext/ED051737.pdf

Ravitch, D. (1990). Education in the 1980's: A Concern for 'Quality'. *Education Week*. https://www.edweek.org/policy-politics/opinion-education-in-the-1980s-a-concern-for-quality/1990/01

Registrar, O. of the U. (n.d.). *Bulletin Archives | Office of the University Registrar*. Retrieved November 11, 2025, from https://registrar.duke.edu/bulletins/bulletin-archives/

Robins, L. S., Fantone, J. C., Oh, M. S., Alexander, G. L., Shlafer, M., & Davis, W. K. (1995). The effect of pass/fail grading and weekly quizzes on first-year students' performances and satisfaction. *Academic Medicine: Journal of the Association of American Medical Colleges*, *70*(4), 327–329. https://doi.org/10.1097/00001888-199504000-00019

Rodríguez-Planas, N. (2022). COVID-19, college academic performance, and the flexible grading policy: A longitudinal analysis. *Journal of Public Economics*, *207*. https://doi.org/10.1016/j.jpubeco.2022.104606

Rojstaczer, S., & Healy, C. (2012). Where a is Ordinary: The Evolution of American College and University Grading, 1940-2009. *Teachers College Record*, *114*(7). https://doi.org/10.1177/016146811211400707

Rutter, M., Maughan, B., Mortimore, P., & Ouston, J. (1979). *Fifteen thousand hours: Secondary schools and their effects on children*. Harvard University Press. https://www.

hup.harvard.edu/books/9780674300262

Sabot, R., & Wakeman-Linn, J. (1991). Grade Inflation and Course Choice. *Journal of Economic Perspectives*, *5*(1), 159–170. https://doi.org/10.1257/jep.5.1.159

Sacerdote, B. (2012). When the Saints Go Marching Out: Long-Term Outcomes for Student Evacuees from Hurricanes Katrina and Rita. *American Economic Journal: Applied Economics*, *4*(1), 109–135. https://doi.org/10.1257/app.4.1.109

Santomauro, D. F., Mantilla Herrera, A. M., Shadid, J., Zheng, P., Ashbaugh, C., Pigott, D. M., Abbafati, C., Adolph, C., Amlag, J. O., Aravkin, A. Y., Bang-Jensen, B. L., Bertolacci, G. J., Bloom, S. S., Castellano, R., Castro, E., Chakrabarti, S., Chattopadhyay, J., Cogen, R. M., Collins, J. K., … Ferrari, A. J. (2021). Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the COVID-19 pandemic. *The Lancet*, *398*(10312). https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(21)02143-7/fulltext

Scherer, Z., & King, M. D. (2025). *Income Gap Between Householders With College Degrees and Those With High School Degrees but No College Widened Over Last Two Decades*. https://www.census.gov/library/stories/2025/09/education-and-income.html

Schneider, J., & Hutt, E. (2013). Making the grade: A history of the A–F marking scheme. *Journal of Curriculum Studies*, *46*(2), 201–224. https://doi.org/10.1080/00220272.2013.790480

Schudson, M. S. (1972). *Harvard Educational Review*.

Sgan, M. R. (1969). The First Year of Pass-Fail at Brandeis University: A Report. *The Journal of Higher Education*, *40*(2), 135–144. https://doi.org/10.1080/00221546.1969.11773370

Shapiro, S. L., Shapiro, D. E., & Schwartz, G. E. (2000). Stress management in medical education: A review of the literature. *Academic Medicine: Journal of the Association of American Medical Colleges*, *75*(7), 748–759. https://doi.org/10.1097/00001888-200007000-00023

Shaw, J. A., Applegate, B., & Schorr, C. (1996). Twenty-One—Month Follow-up Study of School-Age Children Exposed to Hurricane Andrew. *Journal of the American Academy of Child & Adolescent Psychiatry*, *35*(3), 359–364. https://doi.org/10.1097/00004583-199603000-00018

Smallwood, M. L. (1935). *An Historical Study of Examinations and Grading Systems in Early American Universities: A Critical Study of the Original Records of Harvard, William and Mary, Yale, Mount Holyoke, and Michigan from Their Founding to 1900*. Harvard University Press.

Spring, L., Robillard, D., Gehlbach, L., & Moore Simas, T. A. (2011). Impact of pass/fail grading on medical students' well-being and academic outcomes. *Medical Education*, *45*(9). https://doi.org/10.4065/80.12.1613

Stallings, W. M., & Leslie, E. K. (1970). Student attitudes towards grades and grading. *Improving College and University Teaching*, *18*(1), 66–68. https://files.eric.ed.gov/fulltext/ED060054.pdf

Starch, D. (1913). Reliability and Distribution of Grades. *Science*, *38*(983), 630–636. https://doi.org/10.1126/science.38.983.630

Stiglitz, J. E. (1975). The Theory of "Screening," Education, and the Distribution of Income.

The American Economic Review, *65*(3), 283–300. https://www.jstor.org/stable/1804834

Stuart, S. (2019). Reflections on a Decade Leading a Medical Student Well-Being Initiative. *Academic Medicine: Journal of the Association of American Medical Colleges*, *94*(6), 771–774. https://doi.org/10.1097/ACM.0000000000002540

Sullivan, A. L., & Simonson, G. R. (2016). A Systematic Review of School-Based Social-Emotional Interventions for Refugee and War-Traumatized Youth. *Review of Educational Research*, *86*(2). https://doi.org/10.3102/0034654315609419

Sumner, R. G. (1935). What Price Marks? *Junior-Senior High School Clearing House*, *9*(6), 340–344. https://www.jstor.org/stable/30174436

Tardiff, K. (1980). The effect of pass-fail on the selection and performance of residents. *Journal of Medical Education*, *55*(8), 656–661. https://doi.org/10.1097/00001888-198008000-00002

Thorisdottir, I. E., Asgeirsdottir, B. B., Kristjansson, A. L., Valdimarsdottir, H. B., Jonsdottir Tolgyes, E. M., Sigfusson, J., Allegrante, J. P., Sigfusdottir, I. D., & Halldorsdottir, T. (2021). Depressive symptoms, mental wellbeing, and substance use among adolescents before and during the COVID-19 pandemic in Iceland: A longitudinal, population-based study. *The Lancet*, *8*(8). https://www.thelancet.com/journals/lanpsy/article/PIIS2215-0366(21)00156-5/fulltext

Townsley, M., & Kunnath, J. (2022). Exploring State Department of Education Grading Guidance during COVID-19: A Model for Future Emergency Remote Learning. *Education Policy Analysis Archives*, *30*(163). https://eric.ed.gov/?id=EJ1374189

U. S. Department of Education, O. of E. R., & Improvement. (1986). *What works: Research about teaching and learning*. U.S. Government Printing Office. https://files.eric.ed.gov/fulltext/ED263299.pdf

Unemployment rates for people 25 years and older by educational attainment. (n.d.). U.S. Bureau of Labor Statistics. Retrieved November 11, 2025, from https://www.bls.gov/charts/employment-situation/unemployment-rates-for-persons-25-years-and-older-by-educational-attainment.htm

United States Senate: Health, E. L., & Committee, P. (n.d.). *Executive Summary*. Retrieved November 11, 2025, from https://www.help.senate.gov/imo/media/for_profit_report/ExecutiveSummary.pdf

United States Senate: Health, E. L., & Committee, P. (2012). *For Profit Higher Education: The Failure to Safeguard the Federal Investment and Ensure Student Success*. https://www.help.senate.gov/imo/media/for_profit_report/PartI.pdf

U.S. Confidence in Higher Education Now Closely Divided. (n.d.). Gallup. Retrieved November 11, 2025, from https://news.gallup.com/poll/646880/confidence-higher-education-closely-divided.aspx

Uttl, B., White, C. A., & Gonzalez, D. W. (2017). Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching ratings and student learning are not related. *Studies in Educational Evaluation*, *54*, 22–42. https://doi.org/10.1016/j.stueduc.2016.08.007

Valentine, J. A. (1987). *The College Board and the School Curriculum. A History of the College Board's Influence on the Substance and Standards of American Education, 1900-1980*. College Board Publications, Box 886, New York, NY 10101. https://eric.ed.gov/

?id=ED285443

Valsan, C., & Sproule, R. (2008). The Invisible Hands behind the Student Evaluation of Teaching: The Rise of the New Managerial Elite in the Governance of Higher Education. *Journal of Economic Issues*, *42*(4), 939–958. https://doi.org/10.1080/00213624.2008.11507197

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, *45*(3), 1–67. https://doi.org/10.18637/jss.v045.i03

Vosti, K. L., & Jacobs, C. D. (1999). Outcome measurement in postgraduate year one of graduates from a medical school with a pass/fail grading system. *Academic Medicine: Journal of the Association of American Medical Colleges*, *74*(5), 547–549. https://doi.org/10.1097/00001888-199905000-00023

Wachtel, H. K. (1998). Student Evaluation of College Teaching Effectiveness: A brief review. *Assessment & Evaluation in Higher Education*, *23*(2), 191–212. https://doi.org/10.1080/0260293980230207

Weber, G. (1971). *Inner-City Children Can Be Taught to Read: Four Successful Schools. CBE Occasional Papers, Number 18* (18). Council for Basic Education. https://eric.ed.gov/?id=ED057125

Wechsler, H. S. (1977). *The Qualified Student. A History of Selective College Admission in America*. Wiley-Interscience, 605 Third Avenue, New York, NY 10016.

Weems, J. E., Clements, W. H., Quann, C. J., Smith, K., & Schefelbein, B. E. (1971). Pass–fail: Were the hypotheses valid? *College and University*, *46*, 535–556.

Weller, L. D. (1983). The Grading Nemesis: An Historical Overview and a Current Look at Pass/Fail Grading. *Journal of Research and Development in Education*, *17*(1), 39–45. https://eric.ed.gov/?id=EJ288937

Westrick, P. A., Le, H., Robbins, S. B., Radunzel, J. M. R., & Schmidt, F. L. (2015). College Performance and Retention: A Meta-Analysis of the Predictive Validities of ACT® Scores, High School Grades, and SES. *Educational Assessment*, *20*(1), 23–45. https://doi.org/10.1080/10627197.2015.997614

White, C. B., & Fantone, J. C. (2010). Pass-fail grading: Laying the foundation for self-regulated learning. *Advances in Health Sciences Education: Theory and Practice*, *15*(4). https://doi.org/10.1007/s10459-009-9211-1

Widnall, E., Winstone, L., Plackett, R., Adams, E. A., Haworth, C. M. A., Mars, B., & Kidger, J. (2022). Impact of School and Peer Connectedness on Adolescent Mental Health and Well-Being Outcomes during the COVID-19 Pandemic: A Longitudinal Panel Survey. *International Journal of Environmental Research and Public Health*, *19*(11). https://doi.org/10.3390/ijerph19116768

Yang, H., & Yip, C. S. (2003). *An Economic Theory of Grade Inflation*. https://www.researchgate.net/publication/242244547_An_Economic_Theory_of_Grade_Inflation