Thesis

Sophia Yang

Introduction

In the wake of the COVID-19 pandemic, the Trinity College of Arts and Sciences at Duke University introduced academic policy changes. During the Spring 2020 semester, all courses were switched to Satisfactory/Unsatisfactory (S/U) grading as opposed to the traditional graded (A- F) scale. While this was the only semester where all courses were allowed to be S/U and count for all requirements, other policy changes during the pandemic regarding S/U still remain.

Referencing Duke University Undergraduate Bulletins (Registrar, 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. Some faculty members have noted an increase in student usage of S/U and departments are moving towards assigning courses to be graded (A-F) only. The most recent of several proposed changes has been the approval of a proposal to switch in policy defaults such that instructors have to opt-in for allowing S/U, rather than having to specifically opt-out of allowing S/U (Halper, 2025). With the rollback of most pandemic-era policies such as masking and social distancing long gone, this raises the question of whether the S/U policy should return to its pre-pandemic form.

A growing body of literature has examined the impact of the pandemic on student learning and a plethora of studies have been done on pass/fail grading in post-secondary education (although primarily in medical school). Rather than exploring impacts on student outcomes, this thesis seeks to determine how the changes in flexible grading policy have impacted student behaviors with the goal of better informing educators and administrators.

Background and Context

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 regards 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 of 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).

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 material 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 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.

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 had 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 correlate 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 general 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.

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 but 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" (2012)[1].

[1] https://www.help.senate.gov/imo/media/for_profit_report/PartI.pdf

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 (Senate Committee on Health, Education, Labor, and Pensions 2011)[2]. As a result, the outcomes of students at these institutions has 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 (2012)[3].

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 rising the "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 over saturation 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 from U.S. News beginning in 1983 (*Best College Rankings*, n.d.). A dominant 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 only accelerated in the 21st century. In response to the Great Recession (2007-2009) and mandatory spending programs like Medicaid, between 2008 and 2013 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] https://www.help.senate.gov/imo/media/for_profit_report/ExecutiveSummary.pdf
[3] https://www.ed.gov/sites/ed/files/policy/highered/reg/hearulemaking/2012/notice-proposed-rulemaking-march-14-2014.pdf

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 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 and Sproule 2008)[4]. 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 1980s, SET became commonplace in American higher education (Centra 1993; Wachtel 1998)[5][6]. 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 2010)[7]. Multiple studies have found significant, positive correlations between student evaluations and student grades (Langbein 2008; Ellis 2003)[8][9]. On the other hand, SET rankings are not significantly correlated with actual student learning (Uttl 2017)[10]. An article by Neath 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

[4] 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

[5] https://eric.ed.gov/?id=ED363233

[6] https://doi.org/10.1080/0260293980230207

[7] https://link.springer.com/article/10.1007/s10805-010-9117-9

[8] https://doi.org/10.1016/j.econedurev.2006.12.003

[9] https://doi.org/10.1080/00220670309596626

[10] https://doi.org/10.1016/j.stueduc.2016.08.007

(Neath 1996)[11]. Simultaneously, efforts to cut costs and increase profit margins resulted in a rise of nontenured, adjunct faculty (Bettinger and Long 2010)[12]. 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 higher 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[13]. 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 while over the same time period Grade 12 assessment scores decreased in mathematics and did not significantly change in the sciences[14]. 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 and 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[15]. 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 and Sawyer 2004)[16]. 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)[17]. Overall, there is insufficient evidence to suggest that the quality of students has risen, despite significant increases in GPA. GPA is no longer an

[11] Neath, I. (1996). How to improve your teaching evaluations without improving your teaching. Psychological Reports, 78, 1363-1372. https://doi.org/10.2466/pr0.1996.78.3c.1363

[12] https://scholar.harvard.edu/files/btl/files/bettinger_long_2010_does_cheaper_mean_better_-_impact_of_using_adjuncts_-_restat.pdf

[13] https://www.gradeinflation.com/tcr2010grading.pdf

[14] https://www.nationsreportcard.gov/hstsreport/

[15] https://www.proquest.com/docview/225613390?sourcetype=Scholarly%20Journals

[16] https://www.proquest.com/scholarly-journals/is-high-school-gpa-better-than-admission-test/docview/225613390/se-2

[17] 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

effective tool for differentiating skill.

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 (Chan, Hao, & Suen, 2007; Ehlers & Schwager, 2016)[18][19]. For students, they are deprived of feedback and left unmotivated by lack of recognition of exceptional effort (O'Halloran & Gordon, 2014)[20]. For society, grade inflation means graduates are unprepared for the workforce without the skills, dedication, knowledge, and work ethic desired by employers (Franz, 2010; Love & Kotchen, 2010; Yang & Yip, 2003)[21][22][23].

With more degree holders and higher GPAs, degrees are no longer a sufficient edge over other job seekers 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"[24]. 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 (Durk and Perry 2020)[25].

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[26]. 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[27]. 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[28].

[18] https://doi.org/10.1111/j.1468-2354.2007.00454.x
[19] https://econpapers.repec.org/RePEc:oup:cesifo:v:62:y:2016:i:3:p:506-521
[20] https://link.springer.com/article/10.1007/s10734-014-9758-5
[21] https://doi.org/10.1016/j.econedurev.2009.10.013
[22] https://ideas.repec.org/a/pal/easeco/v36y2010i2p151-163.html
[23] https://www.asc.ohio-state.edu/yang.1041/grade-inflation.pdf
[24] https://www.forbes.com/sites/camilomaldonado/2018/07/24/price-of-college-increasing-almost-8-times-faster-than-wages/?utm_source=chatgpt.com
[25] https://eric.ed.gov/?id=ED610721
[26] https://news.gallup.com/poll/646880/confidence-higher-education-closely-divided.aspx
[27] https://www.bls.gov/charts/employment-situation/unemployment-rates-for-persons-25-years-and-older-by-educational-attainment.htm
[28] https://www.census.gov/library/stories/2025/09/education-and-income.html

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[29][30]. 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)[31]. Student satisfaction was measured and found to have increased in two of the papers[32][33]. 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 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 significantly higher than the pass/fail cut-off in[34]. 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[35]. Surveys of students and directors also showed majority belief that pass/fail evaluation hindered the ability to compete for residency (29, 14). Although, other studies have shown that some students believe grades are already arbitrary while others prefer grading for motivation[36]. However, it appears that the actual impact of grading tiers may depend on the institution giving the grades as in practice 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[37]. There was also evidence that pass/fail grading reduced competition and external

[29] https://pubmed.ncbi.nlm.nih.gov/10926029/
[30] https://www.sciencedirect.com/science/article/pii/S0025619611610574
[31] https://doi.org/10.1111/j.1365-2923.2011.03989.x
[32] https://journals.lww.com/academicmedicine/fulltext/2009/05000/a_change_to_pass_fail_grading_in_the_first_two.28.aspx
[33] https://journals.lww.com/academicmedicine/abstract/1995/04000/the_effect_of_pass_fail_grading_and_weekly_quizzes.19.aspx
[34] https://pubmed.ncbi.nlm.nih.gov/7718068/
[35] https://pubmed.ncbi.nlm.nih.gov/7401142/
[36] https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0240587
[37] https://pubmed.ncbi.nlm.nih.gov/10353289/

motivation for grades without decreasing the amount of time students spent studying, defying expectations of increased laziness[38].

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 (Slavin 2019)[39]. Additional studies on the differences between pass/fail and graded students found no consistent difference between student cohorts[40]. Interestingly, some have also suggested the usage of pass/fail as a method of combating grade inflation (Blum 2017)[41] while others argue pass/fail is a cause of grade inflation[42].

In summary, there has been an increased interest in pass/fail arising from a prevalent attitude that educational reforms were 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 and Fantone 2009)[43]. 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 about whether or not pass/fail is a better system than traditional A-F tiered grading.

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[44]. There is a plenitude of literature on crises such as school shootings[45], political conflicts[46][47], and natural disasters[48][49][50]. However, the

[38] https://pubmed.ncbi.nlm.nih.gov/5548587/
[39] https://journals.lww.com/academicmedicine/fulltext/2019/06000/reflections_on_a_decade_leading_a_medical_student.27.aspx
[40] https://europepmc.org/article/med/30392003
[41] https://core.ac.uk/download/pdf/132324597.pdf
[42] https://www.jstor.org/stable/27797025
[43] https://pubmed.ncbi.nlm.nih.gov/20012686/
[44] https://www.tandfonline.com/doi/abs/10.1080/02643944.2014.880123
[45] https://journals.sagepub.com/doi/full/10.3102/0162373715590683?casa_token=MLQZHaa7CC0AAAAA%3AnmMeAqWwjR5xoLHvw1Pf_HpCYkNPlN0WBcPF6sNBUXuPynJzsv3wYds0Wd4Ca5gFJZCQE66oPWRT
[46] https://academic.oup.com/jeea/article/17/5/1502/5292664
[47] https://journals.sagepub.com/doi/full/10.3102/0034654315609419?casa_token=S7N8UD6MVlMAAAAA%3AEAJI_7_3siHstlFQfw-laLYSxgUwJUEeVmI_yLS5VGCdTi743zFWVnaTrIhZZwoQfTy7srTJblOF
[48] https://www.sciencedirect.com/science/article/pii/S0890856709634675
[49] https://www.aeaweb.org/articles?id=10.1257/app.4.1.109
[50] 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-

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[51]. 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[52]. Numerous other studies support disproportionate impact of the pandemic on vulnerable groups[53][54][55]. Yet, grade inflation during this period exceeded multiannual trends and still persists in some academic divisions[56]. In addition to cases of COVID-19, students' physical well-being trended down as there was less engagement in physical activity[57]. Similarly, mental well-being also declined with increased world wide prevalence of depression and anxiety[58]. 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[59][60][61][62].

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 DOE guidance often suggested or even required alternative grading[63]. A review paper on COVID-19 academic changes in higher education identified binary grading as one of their five key themes[64]. It's reported that at least 194 American universities implemented pass-fail policies during the Spring 2020 semester[65]. However, there is a dearth of research specifically evaluating the impact of COVID-19 induced

Louisiana-One-Year-Follow-Up.pdf

[51] https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2021.746289/full
[52] https://journals.sagepub.com/doi/abs/10.3102/0013189X221109178
[53] https://www.pnas.org/doi/full/10.1073/pnas.2020685118
[54] https://www.tandfonline.com/doi/full/10.1080/21681376.2022.2084447
[55] https://www.sciencedirect.com/science/article/pii/S0272775722000103?via%3Dihub
[56] https://www.tandfonline.com/doi/full/10.1080/03075079.2025.2470297?src=recsys
[57] https://jamanetwork.com/journals/jamapediatrics/fullarticle/2794075
[58] https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(21)02143-7/fulltext
[59] https://www.thelancet.com/journals/lanpsy/article/PIIS2215-0366(21)00156-5/fulltext
[60] https://www.academicpedsjnl.net/article/S1876-2859(24)00021-4/fulltext
[61] https://www.mdpi.com/1660-4601/19/11/6768
[62] https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2786919
[63] https://eric.ed.gov/?id=EJ1374189
[64] https://www.tandfonline.com/doi/abs/10.1080/02602938.2022.2140780
[65] https://oudigitools.blogspot.com/2020/03/feedback-alternate-grading-in-crisis.html

flexible grading. In one study using data from Queens College suggested flexible grading policies helped reduce negative impacts of the pandemic, particularly for lower income students[66]. 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, freshman students are more likely to choose flexible grading, and STEM students were less likely to use flexible grading[67]. 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[68]. While there is no question flexible grading changed during the pandemic, its long term effects are unknown.

While a couple 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 in the height of the pandemic, it is unclear whether or not they persisted as education returned to pre-pandemic trends 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 impacts of pandemic policies on flexible grading on highly selective, prestigious, expensive, or private universities.

Purpose of this Thesis

As students during the pandemic have begun and continue to graduate and enter the workforce, it is important to understand how pandemic policies have impacted them. For administrators, it is time to evaluate if changes in policies continue to be supportive in a post-pandemic world. With more disasters possible in the future, it's critical to understand the impact of the pandemic on settings including that of higher education. Hence, this thesis aims to determine whether or not there is evidence to suggest a change in student behavior under persisting pandemic changes. If so, I hope to identify patterns of change.

[66] https://www.sciencedirect.com/science/article/pii/S0047272722000081?casa_token=vIy0DAM8D2MAAAAA:6xlm4a-PRCdeIUgQRzhnJWOnl69rgiO1T2vfUg9EdYaYK_2wpHSAdMIoGpbym3qbOVrkjH6Q

[67] https://pmc.ncbi.nlm.nih.gov/articles/PMC10199666/

[68] https://www.tandfonline.com/doi/full/10.1080/03075079.2025.2470297?src=#d1e327

Data

Institution of Study

Duke is a private, non-profit university located in Durham, NC. It is a highly prestigious and selective with an overall acceptance rate of 4.8% for the most recent Class of 2029[69]. Duke University has roots beginning in the 19th century, but was officially established in 1924[70]. 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 in the mid-1950s to the late 1960s. 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 which allows students to S/U before major declaration (P/F and S/U thresholds are identical).

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[71]. 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 been systematically loosened. In light of this historical context, I characterize 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 the 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

[69] https://today.duke.edu/2025/03/duke-welcomes-newest-members-class-2029#:~:text=All%20totaled%2C%20Duke%20has%20offered%20admission%20to,Class%20of%202029%2C%E2%80%9D%20Duke%20Provost%20Alec%20D.

[70] https://library.duke.edu/rubenstein/uarchives/history/articles/narrative-history

[71] https://coronavirus.duke.edu/updates/page/24/

Knowledge (AOK) are: Arts, Literatures, and Performance (ALP), Civilizations (CZ), Natural Sciences (NS), Quantitative Studies (QS), and Social Sciences (SS)[72]. Undergraduate students matriculating in Fall 2025 and onward (outside the scope of my analysis) follow a new Arts & Sciences Curriculum[73].

Note on Ethics and Data Inaccuracies

All data used in this thesis is 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, there are some aspects that I remain skeptical of but assume are accurate for the sake of this analysis. 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. Results can be seen in Table TODO. While there are some significant discrepancies, I choose to believe the algorithm for academic level must include other factors such as progress towards degree completion. I also noted some questionable 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 relatively low impact on my overall analysis.

[insert num sem vs acad lvl bot table]

Data Cleaning

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 will be included in this thesis. There may be a number of current students who later decide to transfer into Trinity or transfer into 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 other institution. Additionally, my focus is on the "typical" student of which I define as a student who graduates in exactly eight Fall/Spring semesters. For students who matriculated in 2022 and onwards and hence have not had the opportunity to enroll in

[72] https://trinity.duke.edu/undergraduate/academic-policies/curriculum
[73] https://trinity.duke.edu/undergraduate/academic-policies/curriculum-students-enrolling-fall-2025

eight semesters, I assume all students will graduate in eight semesters (although this is not the case).

In addition to dropping the small fraction of courses that were unable to be matched (NAs 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 in 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 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"[74]. 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 semester-student pairs that have 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, etc) 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 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

[74] https://trinity.duke.edu/undergraduate/academic-policies/unsatisfactory-satisfactory-grading-option

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.

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, etc) 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 straight-forward, 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 NAs 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. 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 end of sophomore year and due to how little data there will be for freshman, I cannot possibly extract majors from current freshman and juniors (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 proportion of 200 and above levels courses a student has taken per division to assign major divisions. With a 0.93%, 0.87%, and 0.87% 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. The high threshold for Arts & Humanities logically makes sense, the more classes more likely to be majoring. 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. However, it should also be taken into consideration that 1) many students end up double majoring, 2) students might have minors or certificates in other divisions, 3) students could have other academic interests, 4) courses could be cross-listed across divisions, and 5) categorization of majors is somewhat debatable. To provide as example of 5, as defined by Trinity, African & African American Studies is classified as a Social Science while Asian & Middle Eastern Studies is a Arts & Humanities. Looking at the confusion metrics for each proportion threshold in Figure TODO, it becomes evident that the increase in false positives with higher thresholds offsets the increase in true positives.

Data Engineering

In addition to constructing variables for major divisions and time period relative to COVID-19, I also engineered several other variables as I describe below:

- `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 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

Methodology

Before fitting any models, I split my data into training (80%) and test (20%) sets. I split on student IDs to prevent data leakage. For modeling, I prepare my data with assistance of the `tidymodels` package. To model S/U choice, I create a logistic regression using `glm()` where the response variable is whether or not the course was taken S/U by a particular student. Any tuning of hyperparameters is done with 5-fold cross-validation. Due to the large number of observations, I had some difficulty fitting random effects. Hence, please assume the following sections are do not include this random effect unless specified.

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 roughly 2% of rows were positive for S/U. This means that the overwhelming majority of 98% of rows were taken either 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. Figure TODO shows that class imbalance improves over time with more recent years having a slightly higher proportion of S/Us. Hence, I opted to narrow my dataset to Fall 2015-Spring 2025. This marginally helped with an improved imbalance to 97.7% vs 2.3%, but this is not nearly enough.

Class imbalance is not an unique problem. The literature suggests methods such as the use of weights (source), oversampling of the minority class (source), SMOTE (source), adjusting decision thresholds, and usage of more robust methods such as support vector machines (source). I experimented with several of these techniques and my results are shown below in Table ….

Technique(s)	True Positive	True Negative	False Positive	False Negative
SMOTE	1104	36350	17238	232
Weights	1104	36350	17238	232
Decision Threshold				

Variable Selection

While I believe all variables included have some relevance to student behavior, their impact might be insignificant for my purposes or not generalize well. Techniques used 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. The problem with all-subset selection is its need for 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 also tested nested models on predictors I did not believe would have a significant effect. These included study away status (significant number of NAs), summer course status, and number of academic plans (due to flaws in imputation for post-pandemic students). The likelihood ratio tests I performed supported the inclusion of these variables in my final model.

For co-linear variables, I opted to compare AIC (metric of the predictive accuracy) and BIC (how likely the model is to be the "true" model). I compared the numerical course load a student takes in a given semester to the discrete categorization …….

A summary of all variables considered for modeling and my verdicts are provided in the Appendix.

Final Model

From my preliminary models, the best hyper-parameters I got were a light penalty of 0.0001 on pure LASSO (mixture =1), SMOTE oversampling of the minority class to reach a 1 to 5 ratio of S/U and non-S/U respectively, and the default decision threshold of 0.5. All parameters were testing in combination with each other, except the decision threshold.

After addressing major modeling hurdles, I added a random effect on student ID using `lme4::lmer()`. 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. However, the addition of random effects resulted in a significant increase in run time, so I used a random sample of my observations to be used in the training and test sets.

Model Evaluation

To evaluate model fit I use a variety of metrics including but not limited to:

- Intraclass Correlation Coefficient: $ICC = \frac{\text{random effect variance}}{\text{total variance}}$, how much variation can be explained by random effect
- Dispersion (for general linear mixed model) = $\frac{\chi^2}{df_{residual}}$, measure of the degree of variability in the dataset relative to what is expected by the theoretical model

- Precision = $\frac{\text{TP}}{(\text{TP+FP})}$, how many of the observations predicted as positive are positive
- Recall (aka Sensitivity) = $\frac{\text{TP}}{(\text{TP+FN})}$, true positive rate
- Specificity = $\frac{\text{TN}}{(\text{TP+FN})}$, true negative rate
- F1 = $\frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision+Recall}}$

  *TP, TN, FP, FN: True Positive, True Negative, False Positive, and False Negative, respectively

  *χ^2 is the Pearson chi-square statistic, $df_{residual}$ are the residual degrees of freedom

Results

Exploratory Data Analysis

Before fitting my model, I created several exploratory plots. In these plots, I use a more complete dataset than I do for my model by including mandatory S/U courses, non-traditional students, ….

*mention enrollment at Duke stable across years

[prop of grades given out A, B, etc by AY/term + GPAs]

[prop of overloads, S/U, # plans by CY]

[# S/U required(?) courses]

[prop S/U, S/U allowed courses, course sizes –> across divisions]

[acad lvl and catalog lvl]

[acad lvl/catalog lvl and S/U]

[# overloads and # S/U]

[time periods –> increase S/U decrease in W, 2014 vs 2019 vs 2024]

[student majors/divisions by AY]

Performance Metrics

The confusion matrix from my model is shown in TODO. My model correctly identified 76.4% of negative cases (specificity = 0.764), but it only correctly identified 67.4% of positive cases (sensitivity or recall = 0.674). The precision of my model is 0.99, meaning 99% of all predicted positives were truly positive. Balancing precision and recall, the F1 score was 0.802.

Figure TODO shows the Precision-Recall (PR) curve for my model. The area under the PR curve is an abyssal 0.11. However, due to only 2% of my data being of the positive class, the baseline PR AUC (Area Under the Curve) is 0.02 meaning that my model performs better than a random model.

A plot of the Receiver Operating Characteristics (ROC) curve associated with my model can be seen in Figure TODO. The area under the ROC curve is 0.806, representing good discrimination between positive and negative outcomes. A random model would score 0.5 while a perfect model would score 1.0. Given one positive and one negative case, the model will on average rank the positive one higher 81% of the time.

However, the calibration plot in Figure TODO tells a different story about the fit of my model. My model appears to be poorly calibrated with systematic over confidence. For instance, when the model predicts with an 80% probability of a course being taken S/U, in reality the course is on average taken S/U 10% of the time. At each level of confidence, the model systematically predicts a higher chance of occurrence than supported by reality.

My model has an intra-class correlation (ICC) of 0.96. This means that 96% of the variation in the outcome is due to differences between students. This could mean that ….

Finally, there is an abundance of evidence supporting the presence of under dispersion, meaning that the residual variance is smaller than expected. This could be a result of …

Model Summary

[insert table of exponentiated coefficients and signif]

Discussion

Based on model time period really matters…but not much else

Bc of the high ICC seems like Student level beliefs and stances for and against flexible grading based on cultural beliefs or personal values and whatever else plays a large role in…

S/U was supposed to help with mental health things -> comparing with data from DUCkI it says… so i think …

Limitations

As with most studies, there are numerous limitations with this thesis. To begin, there is systematically missing data with the lack data satisfying my classification of post-pandemic (Fall 2022-Spring 2025). There are certainly errors in my imputation of student major divisions as well as a systematic underestimate in the number of academic plans students graduate with. In many cases, students realize later in their academic careers that they have the time to explore other interests. Additionally, I do not attempt to impute student's minors. 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 missing data from the courses for which location was NA as well as no 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 good solution to this as there is a the 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 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.

Within the dataset, there are also some potentially suspect observations. As mentioned in the Data section, there appear to be some errors in the recorded course enrollments as well as discrepancies between number of semesters and internal academic level calculations. 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 can be displayed.

Other variables not captured in my dataset include other policy changes that were not explicitly denoted in the Duke Bulletins[75] 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. I may also be

[75] https://registrar.duke.edu/bulletins/bulletin-archives/

misinterpreting the language used by the Trinity College. One conscious mistake is 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[76].

Some areas of this thesis are rather arbitrary for sake of simplicity. 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 through the classification of departments into divisions. As mentioned in my methodology, the classification of African & African American Studies as a social science but the classification of Asian & Middle Eastern Studies as an arts & humanities department seems to neglect the interdisciplinary nature of many fields.

Finally, my model seeks to determine student-level behavior rather than course-level behavior. I do not include course IDs in my model and hence 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.

Conclusion

Based on my model, there is clear evidence that students were more likely to … Factors such as …. also impact students' decisions to chose S/U grading… There does not seem to be sufficient evidence that students' major divisions …

Future Work

For the improvement of the work carried out in this thesis, a couple items come into mind. First, it may be worth exploring a hierarchical structure where-in there is some level of variation in student behavior expected at the the division level, the department level, and at the individual level. Other nitpicks include the potential benefit of capturing specific course ID level effects and the use of overall cumulative GPA instead merely looking at previous semester GPAs. 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,

[76] https://trinity.duke.edu/undergraduate/academic-policies/unsatisfactory-satisfactory-grading-option

the shift to the newest Duke undergraduate curriculum beginning in Fall 2025 should also be evaluated for its intersection with student behaviors.

In my thesis, I briefly explored the problem of major prediction based on prior course history. This is a problem that has been explored more in depth by others such as Lang et al. (2022).[77] It could be of interest to identify patterns in major selection at Duke University and compare them with wider shifts across the nation. Similarly, past work has examined differences in grading standards across departments. My thesis was only able to operate on the level of divisions, but a fine-grained internal investigation at the departmental level could be fruitful. An evaluation of how various departments at Duke compare to each other could increase enrollment in less popular departments through calibrating grading scales and workloads to match more attractive courses, or by modifying courses deemed to have disproportionately relaxed standards for their catalog level. A related problem is the inconsistent categorization of majors into divisions. It is understandably hard to categorize all of human knowledge into distinct, disjoint departments. I am curious as to how associated each major, department or institute is interconnected and which are more similar with others in their division and which are more interdisciplinary. One way doing so, I would propose, may 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[78].

Many factors were inaccessible for my analysis such as student demographics, extracurricular, scholarships, and health conditions. Future work could examine correlation between these factors and grading policies to contribute the the growing body of literature on whether or not S/U improves student well-being or increases the amount of time students have for extracurricular commitments. A more specific feature of interest, would be course attributes associated with individual classes. These attributes directly correspond to graduation requirements under Curriculum 2000[79]. At the root, the reason why I was interested in specific academic programs (majors, minors, certificates) was that some programs lack courses that satisfy certain requirements. For instance, I do not believe there has been a course listed in the statistical science department that has had the code for Arts, Literature, and Performance (ALP), at least not for the past several years. 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. Even with the retirement of Curriculum 2000, a summary of the proportion of courses and their associated Areas of Knowledge or Modes of Inquiry listed in each department could help identify interdisciplinary gaps to be filled–a similar initiative could also be done using the new latest curriculum's codes[80].

Beyond flexible grading policies, another common method used to promote academic exploration is the option of auditing courses. From what I could find in the historic Duke bulletins,

[77] https://files.eric.ed.gov/fulltext/EJ1360550.pdf
[78] https://www.sciencedirect.com/topics/computer-science/cosine-similarity
[79] https://trinity.duke.edu/undergraduate/academic-policies/curriculum
[80] https://trinity.duke.edu/curriculum

the Trinity College of Arts and Sciences has been allowing audits in 1960 and then introduced flexible grading with the pass-fail system in 1966[81]. For administrators it may be worth weighing the trade offs and use-cases for flexible grading versus audits. Audits are known for ….pros…. cons…

Code Availability

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

[81] https://registrar.duke.edu/bulletins/bulletin-archives/

References

Appendix

Variable	Data Type	Reason for Interest	Use in Final Model?
term_units	numerical	Students who take more credits may elect to S/U to reduce course load.	No
term_load	categorical	Same as term_units but with less class imbalance.	Yes, better AIC/BIC than term_units
prev_semGPA	numerical	Students who have a higher GPA might be more motivated to maintain it. Students who have a lower GPA might be more motivated to take courses graded to boost GPA.	Yes
num_plans	numerical	Students seeking to satisfy more academic plans could have a greater workload. Alternatively, their additional plans could include courses that satisfy graduation requirements such that they would elect to take more courses graded A-F.	Yes
studied_away	binary	Courses taken as part of Duke Administered programs do count towards student GPA, while Duke Approved programs do not count towards student GPA. In this sense, Duke Approved programs somewhat act as S/U grading.	Yes

Variable	Data Type	Reason for Interest	Use in Final Model?
took_summer_courses	numerical	The pacing, environment, number of courses, and other factors are different during summer terms. Some students may wait to take more challenging classes for during the summer. Summer courses may also be taken to make up for withdrawals.	Yes
is_art_humanity	binary	A student who majors in arts and humanities may be more incentivized to take such courses on a graded basis to count towards major requirements.	Yes
is_social_sci	binary	A student who majors in social sciences may be more incentivized to take such courses on a graded basis to count towards major requirements.	Yes
is_natural_sci	binary	A student who majors in natural sciences may be more incentivized to take such courses on a graded basis to count towards major requirements.	Yes
student_group	ordinal	When a student matriculates may say something about peer pressures and the policies they experienced.	No, multicolinearity with timeperiod and academic_level_bot
timeperiod	ordinal	Policies were different across timeperiods, so student behaviors may have changed.	Yes
term_enrolled	ordinal	Same rationale as timeperiod.	No, colinearity with timeperiod

Variable	Data Type	Reason for Interest	Use in Final Model?
academic_level_bot	ordinal	The amount of importance and effort students have for their courses varies across their academic career. Perhaps "senioritis" results in more S/Us, maybe first-years chose to S/U as they adapt to a new environment, etc.	Yes
num_semesters	numerical	Same rationale as academic_level_bot.	No, colinearity with academic_level_bot
catalog_level	ordinal	It is possible that a greater proportion of higher level courses do not allow S/U or that lower level courses are typically chosen to fulfill general education requirements.	Yes
division	categorical	The academic rigor, requirements, quality of instruction, and other factors placed by different divisions on their courses may vary and influence student decisions.	Yes
num_students	numerical	A larger class size might disincentivize participation and engagement with course materials, resulting in a loss of effort.	
num_overloads	numerical	Similar rationale as term_load but at a student behavior level. If a student has overloaded often in the past, they may be more likely to overload in the future. Alternatively, students may get burnt out or realize later in their careers that they are missing credits.	

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

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

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

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

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.

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

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

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

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

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

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

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 Edu-*

cation. National Commission on Excellence in Education, U.S. Department of Education. https://eric.ed.gov/?id=ED226006

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

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

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.

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

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

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

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/

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

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

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.

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

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

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

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

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