MANAGEMENT SCIENCE
Vol. 00, No. 0, Xxxxx 0000, pp. 000–000
ISSN 0025-1909 | EISSN 1526-5501 | 00 | 0000 | 0001

INFORMS
DOI 10.1287/xxxx.0000.0000

# 30 Million Canvas Grading Records Reveal Widespread Sequential Bias and System-Induced Surname Initial Disparity

Jiaxin Pei
School of Information, University of Michigan

Zhihan (Helen) Wang, Jun Li
Ross School of Business, University of Michigan, {pedropei, helwang, junwli}@umich.edu

The widespread adoption of learning management systems in education institutions has yielded numerous benefits for teaching staff but also introduced the risk of unequal treatment towards students. We present an analysis of over 30 million Canvas grading records from a large public university, revealing a significant bias in sequential grading tasks. We find that assignments graded later in the sequence tend to (1) receive lower scores, (2) receive comments that are notably more negative and less polite, and (3) exhibit lower grading quality measured by post-grade complaints from students.

Furthermore, we show that the system design of Canvas, which pre-orders submissions by student surnames, transforms the sequential bias into a significant disadvantage for students with alphabetically lower-ranked surname initials. These students consistently receive lower scores, more negative and impolite comments, and raise more post-grade complaints because of their disadvantaged position in the grading sequence. This surname initial disparity is observed across a wide range of subjects. For platforms and education institutions, the system-induced surname grading disparity can be mitigated by randomizing student submissions in grading tasks.

*Key words*: behavioral bias, education inequality, system design, textual analysis

## 1. Introduction

Education technology (EdTech) has been reshaping education systems on multiple fronts. In the post-pandemic era, an increasing amount of lectures, assignments, assessments, and teacher-student interactions are streamed online, making learning management systems (LMSs) a critical component of education and learning in schools. LMSs like Canvas, Blackboard Learning, and Google Classroom allow students to easily download learning materials, access open discussions, submit assignments, and receive feedback from teaching staff using a centralized platform.

Although LMSs can improve the efficiency of student-teacher interaction, the unified platform also faces potential risks for education inequality similar to other large digital platforms and computer systems as any subtle issues in system design could affect hundreds of millions of users around the world. Existing studies in computer science, human-computer interaction, and fairness have long documented inequality and biases created by deficit system design or algorithms (Friedman and Nissenbaum 1996). For example, commercial face recognition systems fail to recognize black

**Table 1** **Popular LMSs and online learning platforms all sort student based on their surnames when grading.**

| Type | System | Learner Enrollment | Surname order grading |
|---|---|---|---|
| Learning management system | Canvas | 25M | ✓ |
| | Blackboard | 6.8M | ✓ |
| | Moodle | 8.5M | ✓ |
| | D2L Brightspace | 8.2M | ✓ |
| Online Learning Platform | Coursera | 16M | ✓ |

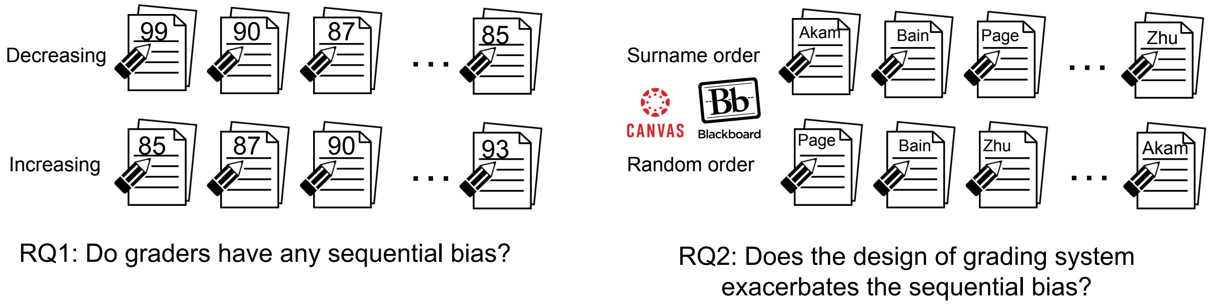
Note: LMS enrollment data of 2022 is from listedtech.com. Coursera enrollment data is from 2020 Coursera Impact Report.

people's faces due to the imbalance issue of training dataset (Leslie 2020) and black patients need to have more severe symptoms to be recommended for the same level of care due to the internal algorithm bias in the healthcare systems (Obermeyer et al. 2019). Besides flawed designs or biased algorithms, researchers also found that some seemingly correct designs may actually lead to unintentionally bad outcomes. For example, (Lambrecht and Tucker 2019) found that a gender-neutral job advertisement recommendation algorithm may still deliver fewer STEM jobs to women simply because showing advertisements to women is more expensive. Given the importance of education to individual career development and social equality, understanding potential biases in education technology platforms holds much significance.

Grading is central to the education process from kindergartens to universities. Grades and teacher comments not only serve as important sources for students to learn from feedback but also play a vital role in future admissions and career development. Therefore, fair grading is an essential and vital part for creating an equal education system. However, equal grading has been notoriously hard to achieve, no matter in K-12 education or higher education settings. Literature documents grading bias against various characteristics of students, such as gender and race (Protivinsky and Munich 2018, Quinn 2020). Assignment feedback can enhance the engagement and performance of students, no matter delivered in paperback or online, though the effectiveness depends on quality and timeliness Evans (2013), Iraj et al. (2021), Wisniewski et al. (2020). Adding to this, existing studies also show that unequal grading treatment significantly impacts students' long-term academic outcomes and personal achievements (Lavy and Megalokonomou 2019, Camille 2020). Grades in the college not only serve the purpose of diagnosing learning outcomes but also act as an indicator of intrinsic quality in the post-graduation recruiting process and board social evaluation contexts.

As a classic sequential task, grading requires a grader to read, examine, and score several to a few hundred student submissions in a sequential manner. In sequential tasks, human beings show widespread behavioral biases in various settings including hiring (Vives et al. 2021), judging (Simonsohn and Gino 2013), and speed dating (Bhargava and Fisman 2014). Existing research has

**Figure 1** **In this study, we ask two major research questions: (1) As a classic sequential task, do graders have any sequential biases in grading? and (2) If there are sequential biases in grading, does the sorting algorithm of the student assignments exacerbate the bias?**


Decreasing
99
90
87
85
Increasing
85
87
90
93
RQ1: Do graders have any sequential bias?
Surname order
Akam
Bain
Page
Zhu
CANVAS
Bb
Blackboard
Random order
Page
Bain
Zhu
Akam
RQ2: Does the design of grading system exacerbates the sequential bias?

proposed theories to model the human behavioral changes in sequential ratings such as learning Bavafa and Jonasson (2021b), fatigue Bavafa and Jonasson (2021a), Dai et al. (2015), contrast effect (Bhargava and Fisman 2014), generosity-erosion effect (Vives et al. 2021), narrow bracketing (Simonsohn and Gino 2013), and affective priming (Chang et al. 2017). Each of the above theories is supported by either empirical or experimental evidence. However, in the context of grading, the connection between grading order and grading outcomes remains unclear. More importantly, while grading has been a core task for teachers for hundreds of years, the recent deployment of LMS introduced a new problem: how are the assignments pre-ordered by the system? In SpeedGrader, the grader system embedded in Canvas LMS, assignment submissions are automatically arranged based on students' surname initials, resulting in graders starting with names like *Andersons* and *Clarks* and only after hours of grading do they reach names like *Wangs* and *Zachs*. Table 1 illustrates that surname initial-ordered grading is the default configuration not only in Canvas but also in the four largest online learning management systems, which collectively account for over 90% of the US and Canadian market.

In this study, as illustrated in Figure 1, we focus on answering two research questions: (1) As a sequential task, are there any sequential biases in grading? In an ideal setting, grading outcomes should not be affected by the order of grading. Here, we define sequential bias as the dependency between the grading outcome and the grading order (e.g. later graded assignments consistently receive higher scores). (2) Popular Learning Management Systems pre-order the student submissions and allow graders to easily navigate, review, and grade. If the graders show a sequential pattern of grading, does the design of the grading system exacerbate the bias, or is able to reduce the issue?

To investigate the impact of sequential grading on grading outcomes, we conducted an analysis of 30 million Canvas grading records obtained from a large public university. Canvas, which is the most widely used online Learning Management System (LMS), had been adopted as the primary system by 32% of all higher education institutions in the United States and Canada by the end of 2020 [1]. Students utilize Canvas to submit their assignments, and teachers utilize the Speedgrader grading system to assess these assignments. We run fixed-effect models that use the grading order of assignments to predict a series of grading outcomes including standardized scores, comment sentiment, comment politeness, and student regrade requests. We controlled a wide range of confounding variables such as the student gender, ethnicity, cumulative GPAs as well as a fixed effect for assignment and grader.

Our analysis revealed several significant findings. Firstly, assignments that were graded later in the sequence tended to receive lower scores. Additionally, feedback provided on assignments graded later was more negative and less polite. This bias had direct implications, as students with assignments graded later were more likely to submit regrade requests, and overall, these students received lower final course grades. While behavioral biases in sequential tasks have long been documented in the existing literature (Vives et al. 2021, Bhargava and Fisman 2014), our results unveiled a robust and substantial implicit behavioral bias in the setting of academic grading.

While it is conceivable that such a sequential bias has persisted for years, the risk posed to each student is akin to a random occurrence, as the grading order may not necessarily be associated with individual characteristics. Therefore, each student is likely to be affected by this bias randomly. However, the widespread adoption of LMS fundamentally alters how assignments are graded in educational systems: as all the popular LMSs pre-order student assignments following the surname order, such a design led to a prominent gap in the average grading order for students with different initials. Students with later initials are more likely to be graded later in the sequence. We calculated the average grading order for each student under different grading modes. We found that when student assignments are graded following the surname order, the average grading order for students with later initials like “Z” is 40, while it is 8 for earlier initials like “A”. Given the sequential pattern in grading, will this design lead to an initial related disparity in student scores?

We investigate the effect of different grading modes on the grading outcomes regarding students with different initials. We run fixed effect models comparing the grading outcome of sequential grading with several baseline groups including a quasi-random group and an autograding group. Through our analysis, we discovered that the system design of sorting student assignments by surnames exacerbates the impact of sequential grading bias, particularly for students with alphabetically lower-ranked surname initials, as they consistently find themselves ranked lower in the grading tasks. Consequently, these students experience significant inequitable treatment and systemic inequality. This disparity transcends various subject areas from social science to humanities

[1] Phil Hill & Associates. State Of Higher Ed LMS Market For US And Canada: Year-End 2020 Edition.

and is further aggregated into course-level disparities: students with later initials are more likely to receive lower scores if their assignments are graded using the surname order model.

To further investigate the causal relationship between grading mode and initial bias, we identify a special setting in our dataset where the assignments are graded backward. In these settings, while the assignments are still surnamed-ranked by the system, graders choose to grade the assignments backward from Z to A. Such a setting creates a comparison group close to a natural experiment. We run a similar fixed-effect model comparing the reverse grading group with the random baseline group. We found that the initial bias was also reversed in this setting: students with earlier initials received lower scores than later initials. The result is significant with the control of student characteristics, course characteristics, and the grader-assignment fixed effect. Such a "natural" comparison provides further evidence that the system design of the LMS converts individual-level grading bias into a widespread surname bias against students with later initials.

While we show a robust and sizable trend that later-graded assignments and later initial students tend to receive lower scores, it is still unclear whether the grading quality is decreasing or whether the graders are simply more accurate at identifying mistakes. To investigate the change of grading quality, we created two measures using student comments submitted after the grades are released: (1) student questions: student comments on Canvas which contain questions, and (2) regrade requests: student comments that complain about the grades or request regrades. We found that later graded assignments are more likely to receive both student questions and regrade requests. Furthermore, later initial students are more likely to submit questions and regrade requests if their assignments are graded in a surname order. Such a result suggests that the quality of grades is decreasing along with the grading process.

Our result has implications for the design of educational platforms as well as the operations of education systems. Through our analysis, we show that the current widely adopted design of the grading system creates unintentional bad outcomes for students with later initials, which could potentially affect over 50 million students around the world. Our result could directly inform the system design of grading systems to pre-order student assignments randomly instead of by surname order. Educational platforms like Canvas and Coursera offer huge convenience and provide new opportunities to students, teachers, and schools. Such a benefit is brought by the power of unified technology platforms where the same system and design are used by hundreds of millions of students around the world. Our results suggest that with all the benefits EdTech has brought, any simple but deficit design of large social systems could potentially create bad outcomes for a large population.

Furthermore, our result contributes to the existing literature on sequential behavioral bias and has implications for education operations. Despite the long debate on class sizes and teacher workload, schools usually face the trade-off of the cost and benefit of increasing the class sizes. Our result characterizes the behavioral bias in sequential grading and suggests that later-graded assignments tend to receive lower scores and the grades have lower quality. Potential solutions include: capping the student amount for each teacher, hiring more graders for large classes, and reducing the overall workload of teachers and graders.

## 2. Literature

Our research contributes to the literature on education inequality, behavioral bias in sequential tasks, name initial bias, and bias in computer systems.

### 2.1. Education Inequality

Education inequality originates from disparity in student capital, including student traits, skills, and family endowments Jackson and Holzman (2020), Quarles et al. (2020). In traditional education contexts, students are treated unequally due to observable identities and characteristics, such as gender Krishna and Orhun (2022), race Redding (2019), cultural background Mason et al. (2014), migrant status Hill and Zhou (2023), and socioeconomic status Gonzalez-Betancor et al. (2021), to name a few. Such unequal treatment comes from both peer interaction Hill and Zhou (2023) and teacher-student interaction Redding (2019), Krishna and Orhun (2022). A recent thread of literature reveals that education inequality widely observed in traditional settings is also present in digital learning environments. For example, gender disparity is still significant in MOOC discussion forums Wang et al. (2023) and virtual classrooms Copur-Gencturk et al. (2022). While existing literature touches on education inequality in the context of education technologies, rare of them have studied how the design of educational technologies could affect students' learning outcomes. Our study examined a seemingly trivial design of the grading system: the pre-ordering of student assignments for grading and directly studied the impact of this design on grading outcomes.

### 2.2. Behavioral Bias in Sequential Tasks

Literature documents that humans deviate from the neutral standard behavioral mode when performing sequential tasks such as food-safety inspections Ibanez and Toffel (2020), ambulance services Bavafa and Jonasson (2021a,b), regular hand hygiene at hospital Dai et al. (2015), to name a few. Theoretically, there are many behavioral mechanisms that may influence human decision-making in sequential tasks. First of all, grading requires expertise where both learning and fatigue could take place Campbell et al. (2015). On the one hand, humans may demonstrate higher efficiency and accuracy as they become more familiar and accumulate more experiences with the task Bavafa and Jonasson (2021b). On the other hand, humans may become fatigued, bored, and impatient when performing complex tasks for a prolonged time, which will impede their performance Bavafa and Jonasson (2021a), Dai et al. (2015). The argument that fatigue can impair self-control in prolonged cognitive tasks is well supported by neuroscience literature Blain et al. (2016). Fatigue is arguably one of the major culprits of substandard performance in sequential tasks. For example, an existing study shows that physicians' empathy for pain assessment tasks is impaired by fatigue and prescribe fewer analgesics during night shifts Choshen-Hillel et al. (2022). Mental and physical fatigue is also prevalent among various occupations such as nurses Barker and Nussbaum (2011), drivers Bavafa and Jonasson (2021a), rail staff Fan and Smith (2020), and students Sievertsen et al. (2016).

Another line of research studying sequential evaluations focuses on the subjective aspect of human ratings. Such studies assume that all human evaluations are relative in nature and therefore, the evaluations of the current item may be affected by previous items (Chang et al. 2017), namely contrast effect (Bhargava and Fisman 2014) or affective priming (Chang et al. 2017). Similarly, the generosity-erosion effect refers to the phenomenon that " as the sequence unfolds, candidates will become more likely to fail if evaluators have previously acted generously" (Vives et al. 2021). (Vives et al. 2021) found that in a high-stake hiring process, candidates after another generously rated candidate are less likely to pass. Such an effect can also happen when the items are evaluated in different sections, namely the narrow bracketing effect (Simonsohn and Gino 2013).

Despite all the existing studies on sequential tasks, it is still unclear whether systematic sequential trend exists in assignment grading. (Wang and Pananjady 2022) conducted crowdsourcing experiment and found that the connections between position and final ratings are not consistent across different items. While (Goldbach et al. 2022) directly studies the sequential dependencies in grading, they only focus on the effect of extreme grades on following scores and therefore excluded early exams and only studied exams with an order higher than 50. Moreover, most of the existing studies use relatively small samples, which can be sensitive to individual variations of graders and assignments.

In this study, we leverage 30 million canvas grading records to study whether sequential patterns exist in assignment gradings. Benefiting from the scale of this dataset, we are able to control the individual and assignment level variations in grading and reveal the sequential pattern of human grading.

### 2.3. Initial Bias

Surname-based disparity generally exists in human interaction contexts, especially when it involves competition and evaluation. Existing studies show that surnames are associated with connotations that influence our perceptions of people, thereby influencing socioeconomic outcomes such as income Arai and Thoursie (2009), job application evaluation Gueguen (2017), and hiring Stefanova et al. (2023). In academic settings, literature also shows that surname initial has an impact on the citability of publications Abramo and D'Angelo (2017), Shevlin and Davies (1997), Tregenza (1997), faculty tenuring, and award-winning Einav and Yariv (2006). Because of the alphabetical ordering of authors for academic publications, researchers with earlier surname initials receive more exposure and acknowledgment for their works, thereby leading to academic success. Alphabetical ordering ubiquitously exists in LMSs, such as mailing lists, student records, grouping, and grading systems. Though the initial-based disparity among researchers has been well documented, the impact of alphabetical ordering on the much larger pool of students is yet to be discovered. This study focuses on the novel setting of online grading and reveals the fact that the alphabetical-ordered system design of LMS translates the human sequential bias into a widespread initial disparity among college students.

### 2.4. System Design and Bias

In recent years, the widespread use of machine learning (ML) and artificial intelligence (AI) systems brings the importance of system fairness to researchers' attention Mehrabi et al. (2021). Friedman and Nissenbaum (1996) define bias in system design as "computer systems systematically and unfairly discriminate against certain individuals or groups of individuals in favor of others" Friedman and Nissenbaum (1996). Bias can arise from many sources within an information system including computer tools, algorithms, and random number generation Friedman and Nissenbaum (1996). Literature shows evidence of bias in system applications such as recommender system Chen et al. (2023), online labor marketplaces Hannak et al. (2017), and web interfaces Metaxa-kakavouli et al. (2018), to name a few. Specifically, system-induced unfairness is prevalent in EdTech applications. For example, a recent study suggests that EdTech webpage designs impose male color bias and can create inequality among students of different gender or ages Santos et al. (2022). Literature also documents implicit biases in digital educational tools Pastushenko and Passmore (2023) and EdTech pedagogy Scardamalia and Bereiter (2008), as well as ethical challenges in EdTech-driven personalized learning Regan and Jesse (2019). In the post-pandemic era, EdTech applications are playing an increasingly instrumental role in post-secondary education. While EdTech brings many benefits to society, a small but hard-to-detect flaw in system design can affect millions of students and create critical social outcomes. Understanding potential biases in EdTech can help foster an equal and inclusive learning environment.

## 3. Background and Data

The objectives of our study are two-folds: (1) to examine the relationship between the grading outcomes and the grading order of student submissions; (2) to understand whether the default grading mode on Canvas transforms the sequential bias of graders into a systematic disparity among students with difference surname initials. In this section, we first describe the context of our study - Canvas learning platform and the integrated grading system, Speedgrader. We then introduce the data we use for empirical analysis.

### 3.1. Context of Study

Learning management systems (LMS) are online software systems that allow schools and teachers to create, manage, and deliver learning materials for education (Turnbull et al. 2020). While traditionally all the class materials are managed in paper files, LMS provides the teachers and students a centralized place to manage class-related materials and interact with others. Since the first creation of LMS, the user base of LMS has been steadily growing. The COVID-19 pandemic further boosted the demand for LMS and over 70 million students around the world have become LMS users by 2023. Popular LMS includes Canvas, Blackboard, Moodle and D2L Brightspace. All the LMSs allow teaching staffs to view student assignments, give grades, and leave comments in the system.

Canvas, founded in 2008, has become the most popular learning management system among K-12 and higher education institutions in the United States and Canada. Canvas has over 25 million enrolled learners by 2020. Canvas allows teachers to easily create and manage course-related materials and also provides students a centralized platform for keeping track of the learning process. Speedgrader is the embedded grading system on Canvas, which allows graders to read assignments, discussions, and quizzes in a specific order and give comments and scores. Under the default settings, Speedgrader sorts assignments in ascending order of student surname initials (for same surnames, sort in ascending order of first names). More specifically, if a student's surname start with "A", their submission is typically ranked as the first several assignments and will be graded earlier. On the contrary, students with surnames starts with "Y" or "X" will find their submissions towards the end of the queue. Graders can choose to sort the submissions by submitting time order. However, this option is located in the system settings manu with low visibility. Our data shows that only less than 0.5% of all submissions are sorted by submitting times. Although graders may manually navigate back and forth among submissions, there is no option in the system to automatically randomize the submissions.

### 3.2. Data

We acquire data [2] from a large public university in the United States. Canvas was rolled out among all courses at the university in 2014 Fall. We collect all available historical data of all

[2] This research has received approval from the University's Institutional Review Board (IRB). The IRB has specifically approved the use of identified student data to link multiple datasets essential for the study, as well as for the extraction of student surname initials. Following IRB approval, the researchers entered into a Memorandum of Understanding (MOU) to gain access to the required datasets. The MOU outlines the specific terms the research team must adhere to in order to safeguard the privacy of student data. In line with the stipulations of both the MOU and IRB, student identifiers were promptly destroyed after (1) the databases were linked and (2) surname initials were obtained.

programs, students, and assignments on Canvas from 2014 Fall semester to 2022 Summer semester. We supplement the Canvas data with university registrar data, which contains detailed information about students' background, demographics, and learning trajectories at the university.

**Canvas Data:** Canvas data contains over 90 million assignment grading records from 2014 Fall semester to 2022 Summer semester. We use the subsample of human-graded assignments for our main analysis, which leads to 30 million submission entries from 144,405 students and 58,162 courses. We further clean the data through four steps: (1) We rule out the impact of extremely small or large courses by removing assignments containing fewer than 5 submissions or more than 400 submissions, which collectively account for 1% of the data. (2) We remove assignments containing fewer than 3 unique values of grades (e.g. Only Pass/Fail) to rule out the confounding factors introduced by over-simplified grading tasks. (3) For similar reason, we remove assignments graded too fast, where the average grading time of one submission is less than 5 seconds. (4) We remove the assignments where evaluation is not based on online submissions, such as attendance or class participation records (e.g. assignment titles contains words like "attendance", "engagement", and "participation").

Our main human-graded sample contains 10 million grading records and 30% of them contain comment messages from the grader. Additionally, we use auto-graded assignments as one of the benchmark comparision group in our empirical analysis because grades on these assignments, typically short, multiple-choice or fill-in-the-blank quizzes, are free from graders' behavioral bias, thereby closer to reflect students' true learning outcomes. The auto-graded sample includes 3.8 million submission entries from 112,237 students and 9,816 courses.

**Student Registrar Data:** We access student registrar data from the learning analytics data architecture (LARC) at the university. The dataset covers all students ever enrolled at the university from 1970 fall to 2022 summer, with a total number of 392,578 students. From this data, we obtain the following variables which we use as controls in our empirical analysis: (1) Students' identifiers including unique university ID and real name, which we use to merge the registrar data to the Canvas data. 90% of the Canvas sample is successfully matched. Also, we extract students' surname initial letters from their names. The identifiers and are destroyed right after these two steps. (2) Student demographics and background information including gender, ethnicity, nationality, minority status, domestic-international student status, high school GPA, and single-parent family indicator. (3) Student learning trajectory at the university including current term GPA, accumulative GPA, graded and non-graded credit units, and academic level (freshmen, sophomore, etc.).

# 4. Metrics and Model

The goal of our study is to answer two research questions: (1) Is there any sequential behavioral bias in grading? (2) Does the design of the online grading system convert the sequential bias into an initial disparity among students? In the following section, we describe the metrics and empirical models we use to investigate these questions.

## 4.1. Variable Construction

### 4.1.1. *Grading Outcomes*

To study the impact of graders' behavioral bias on assignment evaluation, we focus on a series of outcome variables measuring different aspects of the feedback that students receive from graders. In our data, we directly observe the numeric grade and the textual grading comment. We also extract proxies of grading quality from students' post-grade conversations with graders.

*Numerical scores* For each student submission, the grader provides a numerical score for the overall evaluation of the submission. Given that different assignments may have different evaluation criteria and scales, we normalize all scores within each assignment by subtracting the assignment's average grade and then dividing the standard deviation.

*Textual measures of grader comments* Canvas allows graders to give textual comments to submissions, providing graders the opportunity to communicate more nuanced feedback and suggestions to students. Within our data, 30% of student submissions received textual comments. We use state-of-the-art natural language processing models to create two important measures of textual comments: (1) *sentiment*: the overall sentiment of the comment, and it can be negative, neutral, or positive. We map the labels into numbers as -1, 0, and 1 respectively. The sentiment of the comment represents the overall evaluation of a student submission. The sentiment classifier is based on BERT (Devlin et al. 2018), a widely used pre-trained language model for English texts. The model is further fine-tuned on a generic English corpus with sentiment labels (Seethal 2022). (2) *politeness*: the overall politeness of the grader's comment. The politeness score ranges from 1-5 and higher scores indicate higher degrees of politeness. Politeness is an important social information expressed in language and many teaching principles require teachers to be respectful and polite while providing feedback to students. Similar to sentiment analysis, the politeness regressor is created by fine-tuning the BERT model over two politeness datasets (Danescu-Niculescu-Mizil et al. 2013, Wang and Jurgens 2018). The textual measures allow us to study the grading outcomes beyond numerical scores. Table 2 shows examples of grader comments and the predicted scores.

*Grade Quality* Besides the direct numerical and textural grading feedback, it may be more important to ask whether graders' behavioral bias leads to lower quality of the graders' work. Lower grades do not directly imply worse grading quality - graders may give lower grades when they are better able to identify mistakes where the intrinsic quality of grading is higher. One ideal proxy

**Table 2 Examples for Textual Metrics of Grader and Student Comments**

| Grader Comment | Score | Metric |
|---|---|---|
| You completely messed up with Part(c) | negative (-1) | Sentiment |
| A very cool project! Good luck! | positive (1) | Sentiment |
| You misunderstood the instructions. What are you thinking of? This is bad! | 1.4 | Politeness |
| Hi, really great reflection, thanks! | 4.3 | Politeness |
| **Student Comment** | **Score** | **Metric** |
| The only section I've missed an assignment on from not attending was JavaScript. Please fix the grade. | True | Regrade Request |
| Thank you for your suggestions! | False | Regrade Request |
| Would you like me to submit a PDF of the article even though the assignment has already been graded? | True | Student Question |
| I will revise my essay according to your comments! | False | Student Question |

of grading quality is students' reaction after the grades are released. On Canvas, if students have questions or arguments about the grades after they are released, they can directly message the grader in the Speedgrader interface. Logically, students are more likely to comment about the grade if they find it inaccurate. Also, students will raise more questions afterward if they are not fully convinced by the graders' comments. Therefore, we create two measures of grading quality based on student comments after the grades are posted. Specifically, we use regular expressions to identify (1) *regrade requests*, referring to the student comments with grade-related words like "regrade" and "score", and (2) *student questions*, referring to the student comments with a question mark. We use the indicators of whether students raise at least one such comment, separately, as measures of grading quality. To evaluate the precision of this heuristic measure of post-grade requests, We manually examined a random sample of 50 student comments and the precision of this heuristic method in identifying grading outcomes questioned by students is above 0.8. Table 2 shows examples of regrade requests.

### 4.1.2. *Explanatory Variables*

*Grading order* Grading order is defined as each submissions' position in the queue of grading in Speedgrader. To investigate the sequential behavioral bias of graders, we focus on the position of submissions in the sub-queue assigned to each grader, instead of their overall position in the whole assignment which could be graded by more than one grader. Therefore, if an assignment is divided into several submission groups and assigned to multiple graders, we define grading order for each grader separately. For each submission, We obtain grading order from the timestamp of grades. In the case that graders go back to previous graded submissions and alter the grades, the

**Table 3  Description of Variables**

| Category | Variable | Value Type | Description |
|---|---|---|---|
| Grading Outcome | Standardized Scores | continuous | Grade value subtracting average grade divided by standard error |
| | Comment Sentiment | Integer | Sentiment of the assignment comment |
| | Comment Politeness | Continuous | Politeness of the assignment comment, higher scores mean more polite |
| | Student Question | Binary | Whether a student raises questions after the grade is released. |
| | Regrade Request | Binary | Whether a student complaints about the grades or requests regrading after the grade is released. |
| Explanatory Variables | Grading Order | Categorical [1-10, 11-20, 21-30, 31-40, 41-50, 51-60] | Order of the latest grading timestamp of a submission within a assignment-grader queue. |
| | Grading Mode | Categorical [online random, online surname, offline random, autograd] | Grading mode of a set of assignments graded by a grader. Calculated based on correlations between the surname order and actual grading order. If absolute correlation value > 0.8, the grading mode is surname order. If the absolute correlation value is between -0.2 and 0.2, the grading mode is random order. Autograde assignments are graded automatically by the system. |

updated grade reflects the current-stage behavioral status of the grader rather than the previous stage. Therefore, we define the grading order based on the latest timestamp of updating the grade.

The correlation between grading outcomes and grading orders is not necessarily linear. For example, graders' capability of identifying mistakes may go up for the first several submissions as they getting more familiar with the task, and gradually go down for later submissions when they feel tired or bored. To capture the potential nonlinear correlation between grading outcomes and grading order, we introduce categorical dummy variables of grading orders, such as order 1-10, 11-20, 21-30, etc, where each submission falls into exactly one of the grading order bins.

*Grading mode* The Canvas system orders submissions based on student surnames by default and also allows the graders to sort assignments based on submission time. Though most of the graders follow the default sequence of grading, some graders choose to navigate the assignments through the drop-down menu of submissions or the back and forth buttons in a more random manner. Our key explanatory variable for the surname disparity analysis is grading mode, which is a dummy variable indicating whether the submission is graded under the default ascending order of surname initials. This variable is computed based on the correlation between student surname initial order

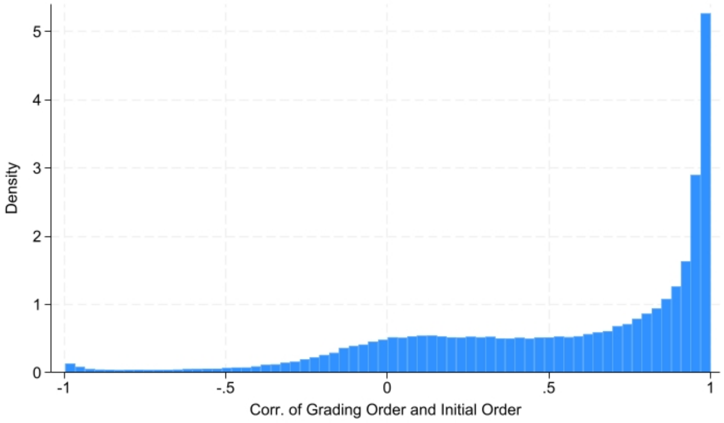
(1 for A, 2 for B, etc.) and grading order. Figure 2 shows the distribution of this correlation for all online-graded submissions. For each assignment-grader submission sequence, If the surname-grading order correlation is above 0.8, we label the grading mode as surname initial order. If the correlation is between -0.2 and 0.2, we label it as the quasi-random grading order. Accordingly, if the correlation between submission order and grading order is over 0.8, we label the grading mode as submission order.

In our main online grading sample, over 40% of assignment-grader bundles follow the surname initial order mode, and 20% are labeled as quasi-random grading mode. Only less than 0.5% of our sample falls into the submission order mode, therefore we exclude these observations in our main analysis. To further account for the confounding factor that late submissions are likely of lower quality, thereby receiving lower grades, We control for each student's order of submission for each assignment in our empirical analysis.

*Student Characteristics* In our empirical analysis, We control for a rich set of student demographics to rule out various individual-level confounding factors that may correlate with the grading outcomes. Specifically, we control for five groups of student characteristics: (1) student ethnicity (Ewijk 2011), including categorical indicators of ethnicity groups, domestic minority group indicator, and underrepresented minority group indicator; (2) gender indicator (Ravenscroft and Buckless 1992); (3) categorical indicators of student nationality (Lindsey and Crusan 2011); (4) indicator of native English speaker (Carroll 2016); and (5) indicator of single-parent family (Pong 1997). Besides demographics, students' historical education records are also strong predictors of future performance, which should be controlled in our empirical analysis. On the one hand, students who historically get high grades are also more likely to be top performers in the future. On the other hand, students' realized performance is also related to their capacity on each course, which, holding the total capacity constant, is negatively correlated with the aggregate workload of their course portfolio. To account for such confounding factors, we control for each student's (6) high school GPA; (7) current term GPA and cumulative GPA of all previous terms; (8) number of graded and non-graded credit units taken, and (9) academic level in the current semester such as freshman, sophomore, etc, in the current semester.

*Course characteristics* Grading outcomes naturally differ from course to course. Courses with more complex and subjective grading tasks, such as literature, writing, and political science courses, may be more subject to the impact of graders' behavioral bias, compared to the courses with analytical grading tasks, such as math, physics, and engineering. Also, holding the grading tasks constant, the impact of graders' behavioral bias may be smaller in courses that are smaller in size, where graders are less likely to get tired and distracted over the relatively shorter grading time. In our empirical analysis, we use normalized grading outcomes where the average grading outcome

**Figure 2** Distribution of correlation scores between student surnames and grading order.



In our analysis, an assignment is considered as surname-order graded if the correlation between grading order and surname order is larger than 0.8. On the other hand, the assignment is graded in a quasi-random order if the correlation is between -0.2 and 0.2. Over 40% of submissions are graded in surname order, while around 20% of assignments are graded in a quasi-random order.

of each assignment is already subtracted. Besides, we control for two course characteristics: (1) class size, measured by the total number of submissions of each assignment; and (2) class subject indicators, including social science, humanities, engineering, medical, science, and others. In Section 5.7, We further investigate the heterogeneity of graders' sequential bias and its impact on student surname disparity.

#### 4.1.3. *Summary of statistics*

Table 4 shows the summary statistics of outcome and explanatory variables on subsamples divided by students' surname initial ranges. Interestingly, U-Z students' assignments receive the highest average score, and the most positive and polite comments among all initial groups. There is a huge disparity in students' grading order - 66% of Students from the A-E group are graded in order 1-10, while only 25% of students from the U-Z group are graded within such range of order. The share of Asian and other ethnicity students is significantly larger in U-Z group compared to other initial groups. We do not observe significant differences among initial groups in other student and course characteristics, such as gender, language, learning history, and class size.

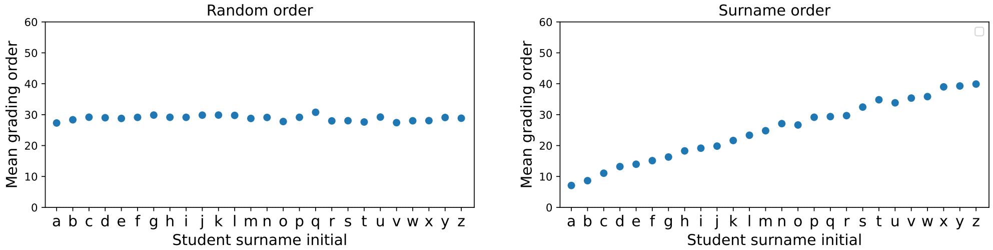
### 4.2. Fixed-Effects Model

#### 4.2.1. *Sequential bias*

Does the grading order affect grading outcomes? To answer this question, we estimate a fixed-effects regression model of the form:

$$y_{iga} = \beta_0 + \beta_1 \text{GradingOrder}_{iga} + \gamma \text{X}_{iga} + FE_g + \epsilon_{iga} \tag{1}$$

**Figure 3 Surname Initial Average Grading Order: By Grading Mode**


Random order

Surname order

Grading order for surname initials when assignments are ranked by students' surname or not. The default grading order leads to large disparities in the initial associated grading order. Around 70% of assignments are graded in this mode.

In the equation above, each observation is an assignment submission. $y_{iga}$ denotes the grading outcome for student $i$ in assignment $a$ graded by grader $g$. We run separate regressions for all the grading outcomes, including standardized scores, comment sentiment, comment politeness, regrade requests, and student post-grade questions. GradingOrder$_{iga}$ includes a set of categorical variables indicating the range of grading order, such as 1-10, 11-20, etc. $X_i$ denotes the set of control variables, including all student and course characteristics we have discussed in Section 4.1.2. We also control for the grader fixed effect to absorb the unobservable characteristics of graders that may influence grading outcomes. To rule out the confounding effect of extremely large or small grading tasks, we use the sample of assignment-graders with at least 5 and at most 60 submissions for this analysis. Robust check results on samples of alternative ranges of submission volumes and samples removing late submissions can be found in Section 5.

### 4.2.2. *System Design*

The second goal of our research is to investigate whether the surname-ordering design of online grading systems transforms the sequential grading bias into a surname disparity of grading outcomes among students. Figure 3 shows the average grading order for each student initial when their assignments are graded in quasi-random or surname order. When assignments are graded randomly, the average grading orders for all initials remain at a similar level. However, if the assignments are graded in the surname order, we observe a large gap in average grading order for students with different initials. For example, the average grading order for initial group "Z" is 40 while it is 8 for initial group 'A". Therefore, to measure the impact of the surname-order grading mode, we use the assignments graded in quasi-random order as the control group. For robustness check, we also run separate analyzes using autograded assignments and offline-graded assignments as alternative control groups.

To formally estimate the impact of surname grading mode and control for confounding factors, we use the fixed-effects regression model of the form:

$$y_{iga} = \alpha_0 + \alpha_1 \text{Initial}_i + \alpha_2 \text{GradeMode}_{ga} + \alpha_3 \text{Initial}_i \times \text{GradeMode}_{ga} + \gamma X_{iga} + FE_g + \epsilon_{iga}, \qquad (2)$$

In the equation above, $y_{iga}$, $X_{iga}$, and $FE_g$ are defined similar to equation (1). $Initial_i$ denotes the student surname initial category of submission $i$, including F-J, K-O, P-T, and U-Z. Note that initial group A-E is omitted due to co-linearity. GradeMode$_{ga}$ is the indicator of surname-order grading mode. $Initial_i \times GradeMode_{ga}$ is the set of interaction terms of initial category indicators and the grading mode indicator.

Our coefficient of interest is $\alpha_3$, which measures the gap of grading outcomes between surname mode and quasi-random mode for each initial category. For example, for initial group U-Z, $\alpha_3^{U-Z}$ denotes the difference between grading outcome under surname mode and that under quasi-random mode, holding all the student and course control variables the same. Similarly, we obtain point estimations of the impact of surname grading mode on each initial category, using quasi-random grading outcomes as the benchmark. Note that compared to the quasi-random mode where all initials have almost the same average grading order, early (late) initial groups have relatively smaller (larger) average grading order. Hypothetically, suppose graders give lower scores as they grade more, we will see positive coefficients for early initial groups like A-E and negative coefficients for late initial groups like P-T and U-Z.

## 5. Results

In this section, we first present the main results for sequential bias (5.1) and system design (5.2). To further demonstrate the relationship between grading mode and surname initial disparity, we show that the the initial disparity is reversed when students are graded in reverse-initial order (5.3). Then we generalize our findings to more grading outcomes, such as final course grades (5.4), sentiment and politeness of grading comments (5.5), and grading quality metrics including student post-grade questions and regrade requests (5.6). Finally, we show the heterogeneity of sequential bias and surname disparity among different subjects (5.7). Due to space limitation, we visualize our key results and provide full estimation results in the Appendix.

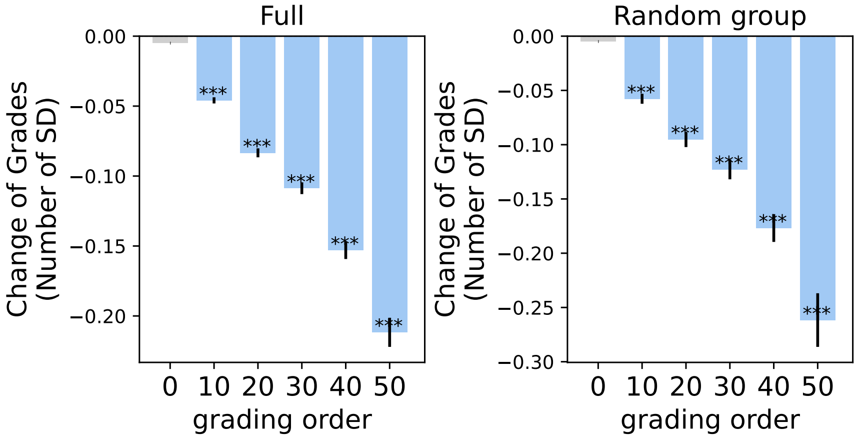
### 5.1. Assignments graded later receive lower scores

Figure 4 presents the coefficients of the grading order variables derived from the regression. Compared with assignments graded in the first session (grading order 0 to 10), assignments graded later in the sequence receive significantly lower scores. More specifically, compared with the first ten assignments, the scores for the 50th to the 60th assignments are lower by 0.2 standard deviations (SD). Such a pattern is also evident in the subsample of assignments graded in random surname

**Table 4 Summary of Statistics**

| Variable | Initial A-E | | Initial F-J | | Initial K-O | | Initial P-T | | Initial U-Z | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Mean | S.D. | Mean | S.D. | Mean | S.D. | Mean | S.D. | Mean | S.D. |
| *Panel A: Grading Outcomes* | | | | | | | | | | |
| Standardized Score | -0.03112 | 0.98921 | -0.01021 | 0.97543 | -0.01361 | 0.97526 | -0.01403 | 0.97725 | 0.01065 | 0.96718 |
| Sentiment | -0.01058 | 0.98722 | 0.00082 | 0.97207 | -0.00183 | 0.96653 | -0.00452 | 0.97073 | 0.00420 | 0.96546 |
| Politeness | -0.00595 | 0.98336 | 0.00095 | 0.96666 | -0.00157 | 0.96395 | -0.00244 | 0.96379 | 0.00267 | 0.96376 |
| Grade Quality - Questions | 0.00000 | 0.03238 | 0.00003 | 0.03311 | 0.00003 | 0.03324 | -0.00006 | 0.03136 | 0.00002 | 0.03504 |
| Grade Quality - Score Comments | -0.00001 | 0.02680 | 0.00003 | 0.02769 | 0.00001 | 0.02740 | -0.00001 | 0.02677 | 0.00002 | 0.02882 |
| *Panel B: Explanatory Variables* | | | | | | | | | | |
| Grading Order 1-10 | 0.66073 | 0.47346 | 0.49993 | 0.50000 | 0.36735 | 0.48208 | 0.28073 | 0.44936 | 0.25154 | 0.43390 |
| Grading Order 11-20 | 0.19601 | 0.39697 | 0.30640 | 0.46100 | 0.35908 | 0.47973 | 0.37842 | 0.48499 | 0.34040 | 0.47384 |
| Grading Order 21-30 | 0.07917 | 0.27000 | 0.11735 | 0.32184 | 0.16362 | 0.36993 | 0.17290 | 0.37816 | 0.18605 | 0.38915 |
| Grading Order 31-40 | 0.04091 | 0.19809 | 0.04777 | 0.21329 | 0.07453 | 0.26264 | 0.10344 | 0.30453 | 0.11819 | 0.32283 |
| Grading Order 41-50 | 0.01798 | 0.13287 | 0.02295 | 0.14973 | 0.02877 | 0.16717 | 0.05304 | 0.22412 | 0.07322 | 0.26051 |
| Grading Order 51-60 | 0.00521 | 0.07197 | 0.00560 | 0.07460 | 0.00665 | 0.08127 | 0.01146 | 0.10645 | 0.03061 | 0.17225 |
| Ethnicity - Black | 0.05476 | 0.22751 | 0.05126 | 0.22054 | 0.04702 | 0.21169 | 0.04059 | 0.19733 | 0.04377 | 0.20459 |
| Ethnicity - Hispanic | 0.06857 | 0.25273 | 0.05550 | 0.22895 | 0.07063 | 0.25620 | 0.07295 | 0.26005 | 0.04674 | 0.21108 |
| Ethnicity - Asian | 0.13700 | 0.34385 | 0.19157 | 0.39353 | 0.20324 | 0.40241 | 0.17431 | 0.37938 | 0.30688 | 0.46120 |
| Ethnicity - Other | 0.13153 | 0.33797 | 0.12835 | 0.33448 | 0.14127 | 0.34830 | 0.12797 | 0.33406 | 0.19351 | 0.39505 |
| Native English Speaker | 0.54232 | 0.49821 | 0.54510 | 0.49796 | 0.52089 | 0.49956 | 0.54276 | 0.49817 | 0.45246 | 0.49774 |
| Female | 0.52762 | 0.49924 | 0.52960 | 0.49912 | 0.52239 | 0.49950 | 0.52366 | 0.49944 | 0.52300 | 0.49947 |
| International Student | 0.07937 | 0.27031 | 0.09296 | 0.29037 | 0.11093 | 0.31405 | 0.08465 | 0.27837 | 0.25092 | 0.43354 |
| Domestic Minority | 0.27455 | 0.44629 | 0.29441 | 0.45578 | 0.31406 | 0.46414 | 0.29166 | 0.45453 | 0.30333 | 0.45970 |
| Underrepresented minority | 0.14346 | 0.35054 | 0.12389 | 0.32946 | 0.13635 | 0.34316 | 0.13096 | 0.33736 | 0.10865 | 0.31120 |
| High School GPA | 2.71052 | 1.75915 | 2.74707 | 1.74227 | 2.68065 | 1.77126 | 2.78805 | 1.72417 | 2.36904 | 1.87524 |
| Single Parent Family | 0.14007 | 0.34706 | 0.13186 | 0.33833 | 0.13903 | 0.34598 | 0.13080 | 0.33718 | 0.12072 | 0.32580 |
| Current GPA | 3.39525 | 0.90505 | 3.41070 | 0.90082 | 3.40012 | 0.91356 | 3.40084 | 0.90577 | 3.44279 | 0.90425 |
| Cumulative GPA | 2.79218 | 1.44940 | 2.80897 | 1.45427 | 2.78465 | 1.46433 | 2.80288 | 1.45059 | 2.79870 | 1.50122 |
| Number of credits - Graded | 12.36800 | 4.47316 | 12.45014 | 4.46165 | 12.38227 | 4.48702 | 12.40902 | 4.46808 | 12.30964 | 4.52466 |
| Number of credits - Non-graded | 1.42147 | 2.92097 | 1.41524 | 2.91022 | 1.42525 | 2.90826 | 1.39886 | 2.90621 | 1.50649 | 2.95799 |
| Class size | 28.51917 | 14.12583 | 28.64818 | 14.15314 | 28.71652 | 14.19587 | 28.62528 | 14.17620 | 29.22752 | 14.20960 |

**Figure 4 Assignment Grades and Grading Order**



Assignments graded later receive lower scores on both the full sample and the random grading order subsample. *** p<0.01, ** p<0.05, * p<0.1.

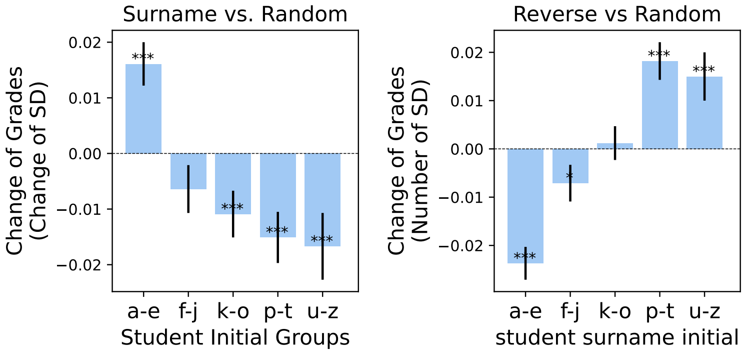
order, suggesting that the impact of grading order on grades is not caused by the natural differences of students with different surnames. While some existing studies found no significant correlations between score and grading order (Bhargava 2008, Goldbach et al. 2022), they use relatively small samples from limited types of classes. Our large-scale analysis based on 10 million human grading records suggests that grading order is negatively correlated with student scores with a relatively large effect size (0.2 SD).

Table A1 in the appendix shows the full regression results for the sequential bias analysis. We observe that both student ethnicity and gender correlate with their assignment grades. Female students receive 0.0814 SD of scores higher than male students, resonating with existing findings that female students are more motivated (Aragon and Johnson 2008) and usually achieve better academic performances (Buchmann and DiPrete 2006). Asian students' grades are higher than white students by 0.0255 SD, while black students receive 0.12 SD of grades lower than White students. Such a difference might be caused by student performance differences and deeper educational inequalities (Troyna 2012). We also found that English native-speaking and international students have higher assignment scores. Regarding student academic performance, We observe that students with higher high school GPAs and higher cumulative GPAs receive higher assignment grades. Furthermore, while existing studies have been debating on whether early-initial students receive better grades (McCullough and McWilliams 2011, Cauley and Zax 2018), we observe that later-initial students receive slightly higher grades than earlier-initial students, with other control variables including ethnicity and student academic performances. Moreover, while we observe statistically significant results for all the above control variables, the effect size of most of the control variables is relatively small (less than 0.15 SD) except for the student's current GPA. However, the effect size for grading order is significantly more prominent compared with other control variables (50th - 60th graded assignments received grades that are 0.2 SD lower than the 1st - 10th graded assignments), suggesting that grading order plays an important role in the grading process.As a robustness check, we remove all the late submissions (bottom 10% sorted by submission time) as late submissions might be associated with lower performances (Sabnis et al. 2022) and still observe a similar result. We observe a similar trend in this experiment, suggesting that the behavioral bias is robust across different settings.

### 5.2. System design leads to surname-initial disparity in scores and rankings.

We analyze human-graded assignments for online submissions and compare the case of surname-initial grading order and quasi-random grading order. Figure 5 shows the coefficient estimations of each initial category interacted with the sequential grading indicator. Students with later surname initials (e.g. K-O and U-Z) receive lower grades when their assignments are graded in the surname order, compared with the quasi-random grading mode. Similarly, such disparity is also larger for later initial categories, where students with initials U-Z tend to receive scores that are 0.15 SD lower than the random grading group. Moreover, earlier initial students benefit from this system design as their assignments are consistently graded earlier in the sequential. Students with initials A-Z

**Figure 5** **System design and Surname Initial Disparity**



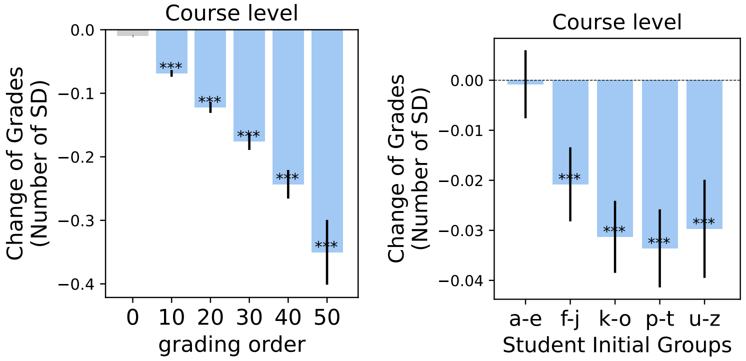
Compared to assignments graded in random order, assignments graded in initial order give early-initial students higher grades and late-initial students lower grades (left). When the assignments are graded backward from Z to A, the initial disparity is also reversed. As students with later initials are graded earlier in this setting, they receive overall higher scores than students with earlier initials (right). *** p<0.01, ** p<0.05, * p<0.1.

tend to receive scores that are 0.16 SD higher than the random grading group. These results suggest that the surname-initial grading mode translates the sequential grading bias into a surname-based disparity in grading outcomes, disproportionally affecting students with lower-ranked surname initials. Table B1 in the appendix shows the detailed regression results. Similar to the sequential bias results, we observe that later initial students tend to perform better than earlier initial students on the random grading sample by a relatively small margin. The coefficient for other control variables generally remains at a similar level as presented in the sequential bias regression table.

### 5.3. Initial disparity is reversed in the reverse grading group

While most of the assignments are graded in surname order, a subsample of assignments is graded backward from Z to A. If the system design of the grading system indeed leads to the initial disparity of student grades, in the reverse grading setting, the initial disparity should also be reversed. Figure 5 (right) shows the result of the reverse grading test. We found that when the assignments are graded backward from Z to A, the initial disparity is also reversed accordingly: compared with the random grading group, students with later initials receive higher scores (around 0.15 SD) when their assignments are graded in the reverse model. While earlier initial students are negatively affected by the reverse grading mode as their assignments are graded later in this setting. Such a result provides stronger causal evidence that the the design of canvas system converts individual-level sequential grading bias into an initial disparity in scores. Furthermore, as most of the assignments are graded following the default setting from A to Z, such a design creates systematic disadvantages for students with later initials.

**Figure 6** **Sequential Bias and Surname Initial Disparity: Course Grades**



(a) Grading order (b) Initial category

(a) At the course level, students graded further behind on average are receiving lower scores. (b) Such a bias is converted into surname-initial disparities. Compared with the quasi-random grading mode, students with later initials receive lower scores if their assignments are graded in surname initial order. *** p<0.01, ** p<0.05, * p<0.1.

### 5.4. Grading bias is associated with surname-initial disparity in final course grades.

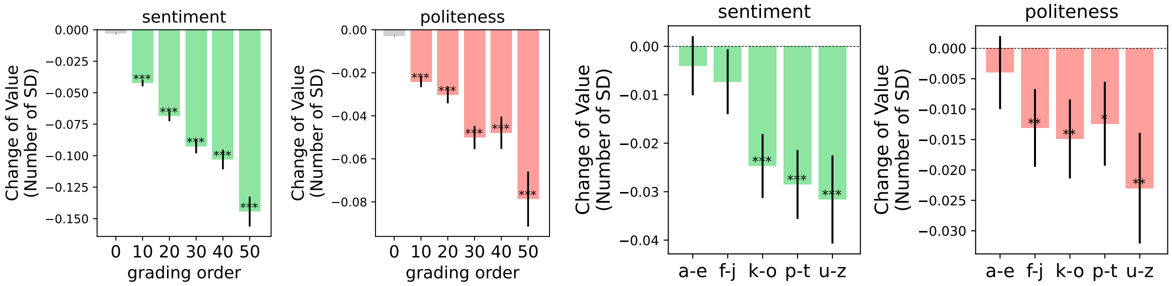
Does the sequential grading bias within assignments accumulate into disparities in final course grades and rankings by surname initial? To answer this question, we analyze course-level outcomes - students' final course grades. The final grade is determined by scores of all graded items on Canvas and the pre-specified weights of each item. As shown in Figure 6a, there exists a similarly sizeable and significant pattern that students graded further behind on average receive lower final scores and are ranked lower in the class.

We further investigate whether the grading mode also translates the grading bias into the surname-initial disparity in final course scores. We use the average correlation between grading order and surname-initial order among all assignments and the same threshold values as the previous section to define surname-initial graded courses and quasi-random graded ones. Figure 6b shows that compared with courses adopting quasi-random grading mode, the courses following surname-initial grading order tend to give students with later surname initials lower final grades and lower rankings in the class. Overall, since the graders and their grading modes within a course are largely fixed, the effect of sequential grading bias accumulates over time and leads to systematic surname-initial disparity in students' final scores, thereby influencing their academic records and potential competitiveness in post-grad job markets.

### 5.5. Grading bias and initial disparity are also evident in assignment comments.

Receiving feedback is an important component of a student's learning process and graders may leave textual comments on students' assignments to provide suggestions and external explanations of the grades. As more and more classes are transferred into online learning mode, assignment

**Figure 7 Sequential Bias and Surname Initial Disparity: Grading Comments**



(a) Grading order

(b) Initial category

(a) Assignments graded later are receiving more negative and less polite comments. (b) Such a bias is converted into initial-related disparities in comments. Compared with the random grading mode, students with later initials are receiving more negative and less positive comments if their assignments are graded in surname initial order. *** p<0.01, ** p<0.05, * p<0.1.

comments become an important channel of teacher-student interaction. Are the sequential grading bias and the resulting initial disparity also reflected in the textual comments?

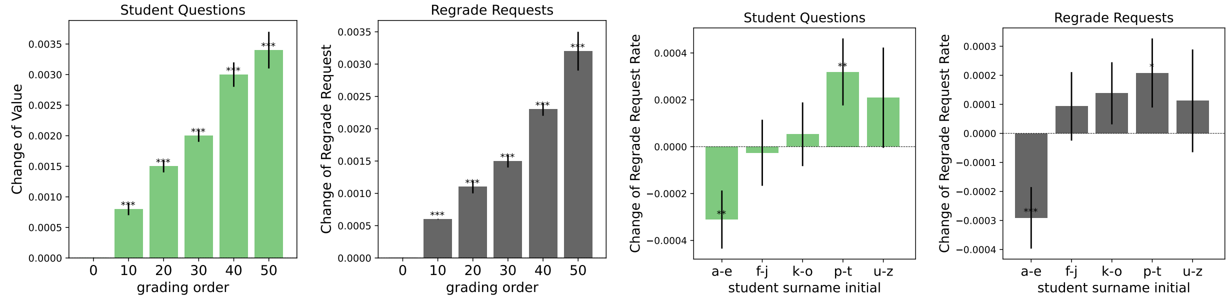
As shown in Figure 7a, assignments graded later tend to receive more negative and less polite comments. Assignments graded in the 50-60th positions receive comments that are 0.13 SD more negative and 0.08 SD less polite than those graded in 1-10th positions. Moreover, such a pattern is further converted into biases against students with later initials as shown in Figure 7b. This result indicates that the sequential grading bias and surname initial disparity are evident in not only the score but also more nuanced characteristics of the comments. One might argue that the sentiment and politeness of the comment could correlate with the overall submission score. In our regression, we controlled for the assignment scores and such a result suggests that the graders might behave more negatively and more impolitely for later assignments and the design of the system converts such a behavioral bias into initial disparities against students with later initials. Table A2 and Table B2 in the Appendix show the full regression results.

### 5.6. Grades have lower quality for later-graded assignments and students with later initials.

Existing studies suggest that people gradually learn to master the task while also getting fatigued and distracted in sequential, repeated task settings (Campbell et al. 2015). While we have shown that score and comment sentiment both significantly decrease with grading order, does the quality of grades also decrease with grading order and for students with later initials?

As shown in Figure 8c, assignment scores receive more post-grade questions from students when the assignments are graded later in the sequence, suggesting that students are generally more unsatisfied with the current grades or the explanation of the grades. Such a pattern also translates

**Figure 8 Grading Quality**



(a) Student Question (b) Regrade Request (c) Student Question (d) Regrade Request

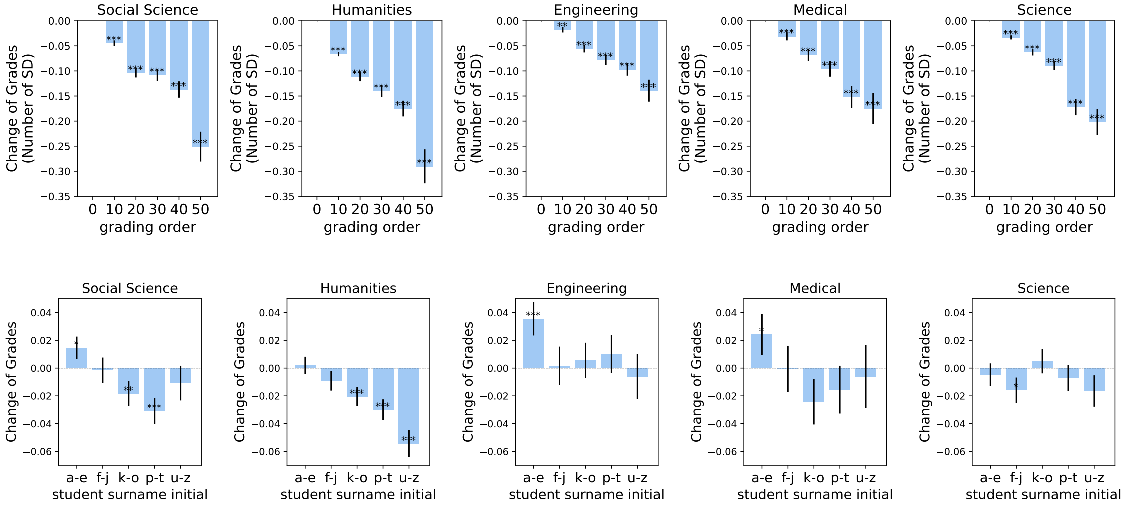
(a) Assignments graded later receive more post-grade questions from students. (b) Students graded later are more like to request regrade. (c) Students with later (earlier) initials are more (less) likely to post questions through comments when graded in surname order. (d) Students with later (earlier) initials are more (less) likely to ask for regrade through comments when graded in surname order. *** p<0.01, ** p<0.05, * p<0.1.

into the disparity by initial groups (Figure 8d). Further, Figure 8a shows the regression coefficient for regrade requests. Regrade request is a stronger signal that students disagree with the points taken off and is directly linked to the quality of the grades. Submissions in later groups are more likely to receive regrade requests from students, suggesting that these scores tend to be less accurate. Such a bias is reinforced by the grading system design (Figure 8b). Compared with assignments graded in a random order, initial-order grading leads to a higher likelihood of regrading requests from students for later initial categories such as *P-T* and *U-Z*. One might argue that students receiving lower scores could naturally be more likely to complain about their grades. Please note that we have controlled for the scores for student assignments and such a result suggests that the final regression coefficients reflect the grading quality beyond student grades. Overall our results on student post-grade questions and regrade requests suggest that graders are becoming less consistent and accurate moving along the grading sequence. The system design of the Canvas system further converts such a behavioral bias into widespread initial disparities in grading quality.

### 5.7. Sequential pattern of grading and initial disparity is evident across subjects

Different subjects may have different distributions of student groups and different types of assignments. Does the sequential trend and initial disparity of grading persist in assignments of different subjects? Figure 9 shows the sequential bias and the surname initial disparities on different subjects. For all the subjects, later-graded assignments receive lower scores, suggesting the sequential bias of grading persists across different subjects in higher educational institutions. For initial disparity, we found that the result is more evident for Social Science and Humanities classes, while the gap for assignments in Engineering and Medical is relatively smaller compared with other subjects.

**Figure 9 Sequential pattern and initial disparity in different subjects**



Sequential grading bias is evident across subjects. Social science and humanity subjects present larger surname disparities while the disparities for engineering, medical, and science subjects are smaller. *** p<0.01, ** p<0.05, * p<0.1.

This is potentially due to the nature of the assignments: social science assignments are usually more subjective and might be harder to grade, while the assignments in engineering and medical could be more objective.

## 6. Discussion

Grading is central to the education process and student performances are also key to their future career development. Receiving fair grading is important for both students as well as the entire society. In this study, by analyzing 30 million grading records, we show how grading order affects the grades students receive: assignments graded later receive lower scores, more negative and less polite comments. Moreover, later grades in a sequence tend to have lower quality and are more likely to receive regrade requests from students. Furthermore, as the Canvas system uses surname order as the default grading mode, the grader-level sequential grading bias is further converted into widespread surname initial disparities in grading: compared with auto-graded and randomly graded assignments, the default grading mode of the Canvas system leads to significantly lower scores for students with later initials. Moreover, students with later initials are also more likely to request regrading after the grading is posted. Such a result has implications for the design of not only education technologies but also digital platforms where sequential task is central to its operations.

### 6.1. Implications for Sequential Tasks

Sequential tasks are ubiquitous in human society: from government employees processing tax documents to workers assembling products in the factory. Existing literature has pointed out many potential behavioral mechanisms behind sequential tasks including learning effect Campbell et al. (2015), Bavafa and Jonasson (2021b), fatigue effect Bavafa and Jonasson (2021a), Dai et al. (2015), Choshen-Hillel et al. (2022), contrast effect (Bhargava and Fisman 2014), generosity-erosion effect (Vives et al. 2021), narrow bracketing (Simonsohn and Gino 2013), and affective priming (Chang et al. 2017). However, grading is a complex scenario where many factors can affect the grading outcomes and none of the existing studies could fully predict whether there will be a sequential trend in grading. Our study leverages the power of a large and rich dataset which contains 30 million human grading records from a large university. With carefully designed analysis, our study reveals a sizeable and robust sequential pattern that later-graded assignments receive lower grades by a large margin. Furthermore, using text analysis methods, we are able to construct two new measures of grading quality: post-grade student questions and regrade requests. These two measures allow us to reveal a potential quality change in the grading process: large-graded assignments receive more student questions and regrade requests. Such a result suggests that the grading quality for later assignments becomes lower. Grading plays a central role in the education system and therefore the managerial implications of our result are significant. Deciding class size is a challenging and tricky task for many school managers. Larger classes mean that there will always be students whose assignments got graded later and our results provide evidence that this might hurt students whose assignments are graded later. Therefore, school managers should consider the negative effect of large class sizes from the perspective of teaching staff's fatigue and consider new policies to reduce their workload.

### 6.2. Implications for System design

Individual-level behavioral bias in sequential tasks is ubiquitous and has long been documented in previous research (Ibanez and Toffel 2020, Bavafa and Jonasson 2021a,b). Traditionally when teachers are grading homework offline, such a behavioral bias seems inevitable – students graded later may receive feedback with lower quality and more biased scores. In the meantime, since the grading is generally down in a random order, it is also likely that such a bias may not create systematic discrimination against certain groups of students. However, with all the benefits and conveniences it brings, education technology platforms like Canvas may create larger and more serious discrimination against certain groups of students. How to sort student assignments for grading may seem to be a trivial and unimportant function, and sorting students based on surname initials or submission time also seems to be reasonable. However, our findings suggest that such a seemingly unimportant function converts individual-level grading bias into large-scale initial-related disparities in grading. Besides Canvas, the most popular learning management systems and education technology platforms like Coursera, Blackboard, and Moodle all sort students based on their surnames. As academic performances could affect the long-term career development of students, such a design could cause widespread harm to students with later initials. Besides the design of computer systems, our result also has implications for "design" in a broader sense. Sorting people by their surnames is widely adopted in many settings from author ranking to your smartphone contacts. Our result suggests that such a seemingly natural and reasonable design could create harm to a large population of people. Managers or system builders should consider the potential harm and biases in future cases when want to adopt a similar setting that ranks people by surnames.

### 6.3. Limitations and Future Research

Our study has the following limitations. First, we use data only from one University and one platform. Therefore, while we show that the effect is consistent across different disciplines, further studies are needed to verify the robustness of the result in different schools. Furthermore, while all the popular LMS platforms adopt a similar design of ranking students by their surnames, we only studied the Canvas platform. Second, while our result on student questions and regrade requests points to the explanation of fatigue leading to decreased grading quality, understanding the nuanced behavioral mechanism behind the sequential bias requires carefully designed lab experiments and we consider it as a future research direction.

## References

Abramo, Giovanni, Ciriaco Andrea D'Angelo. 2017. Does your surname affect the citability of your publications. *Journal of Informetrics* **11**(1) 121–127.

Aragon, Steven R, Elaine S Johnson. 2008. Factors influencing completion and noncompletion of community college online courses. *The Amer. Jrnl. of Distance Education* **22**(3) 146–158.

Arai, Mahmood, Peter Skogman Thoursie. 2009. Renouncing personal names: an empirical examination of surname change and earnings. *Journal of Labor Economics* **27**(1) 127–147.

Barker, Linsey M., Naury A. Nussbaum. 2011. Fatigue, performance and the work environment: a survey of registered nurses. *Journal of Advanced Nursing* **67**(6) 1370–1382.

Bavafa, Hessam, Jonas Oddur Jonasson. 2021a. The distributional impact of fatigue on performance. *Management Science* **Forthcoming**.

Bavafa, Hessam, Jonas Oddur Jonasson. 2021b. The variance learning curve. *Management Science* **67**(5) 3104–3116.

Bhargava, Saurabh. 2008. Perception is relative: Sequential contrasts in the field. Tech. rep., UC Berkeley Working Paper.

Bhargava, Saurabh, Ray Fisman. 2014. Contrast effects in sequential decisions: Evidence from speed dating. *Review of Economics and Statistics* **96**(3) 444–457.

Blain, Bastien, Guillaume Hollard, Mathias Pessiglione. 2016. Neural mechanisms underlying the impact of daylong cognitive work on economic decisions. *Proceedings of the National Academy of Sciences* **113**(25) 6967–6972.

Buchmann, Claudia, Thomas A DiPrete. 2006. The growing female advantage in college completion: The role of family background and academic achievement. *American sociological review* **71**(4) 515–541.

Camille, Terrier. 2020. Boys lag behind: how teachers' gender biases affect student achievement. *Economics of Education Review* **77**(2020) 101981.

Campbell, Danny, Marco Boeri, Edel Doherty, W. George Hutchinson. 2015. Learning, fatigue and preference formation in discrete choice experiments. *Journal of Economic Behavior & Organization* **119** 345–363.

Carroll, Bailey Alison L., Patricia E. 2016. Do decision rules matter? a descriptive study of english language proficiency assessment classifications for english-language learners and native english speakers in fifth grade. *Language Testing* **33**(1) 23–52.

Cauley, Alexander, Jeffrey S Zax. 2018. Alphabetism: the effects of surname initial and the cost of being otherwise undistinguished. *Available at SSRN 3272556* .

Chang, Seah, Chai-Youn Kim, Yang Seok Cho. 2017. Sequential effects in preference decision: Prior preference assimilates current preference. *PloS one* **12**(8) e0182442.

Chen, Jiawei, Hande Dong, Xiang Wang, Fuli Feng, Meng Wang, Xiangnan He. 2023. Bias and debias in recommender system: a survey and future directions. *ACM Transactions on Information Systems* **41**(3) 67.

Choshen-Hillel, Shohan, Ido Sadras, Tom Gordon-Hecker, Shir Genzer, David Rekhtman, Eugene M. Caruso, Koby L. Clements, Adrienne Ohler, David Gozal, Salomon Israel, Anat Perry, Alex Gileles-Hillel. 2022. Physicians prescrbe fewer analgestics during night shifts than day shifts. *Proceedings of the National Academy of Sciences* **119**(27) e2200047119.

Copur-Gencturk, Yasemin, Ian Thacker, Joseph R. Cimpian. 2022. Teacher bias in the virtual classroom. *Computers & Education* **191** 104627.

Dai, Hengchen, Katherine L. Milkman, David A. Hofmann, Bradley R. Staats. 2015. The impact of time at work and time off from work on rule compliance: the case of hand hygiene in health care. *Journal of Applied Psychology* **100**(3) 846–862.

Danescu-Niculescu-Mizil, Cristian, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, Christopher Potts. 2013. A computational approach to politeness with application to social factors. *arXiv preprint:1306.6078* .

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* .

Einav, Liran, Leeat Yariv. 2006. What's in a surname? the effects of surname initials on academic success. *Journal of Economic Perspectives* **20**(1) 175–188.

Evans, Carol. 2013. Making sense of assessment feedback in higher education. *Review of Educational Research* **83**(1) 70–120.

Ewijk, Reyn van. 2011. Same work, lower grade? student ethnicity and teachers' subjective assessments. *Economics of Education Review* **30** 1045–1058.

Fan, Jialin, Andrew P. Smith. 2020. Effects of occupational fatigue on cognitive performance of staff from a train operating company: a field study. *Frontiers in Psychology* **11** 558520.

Friedman, Batya, Helen Nissenbaum. 1996. Bias in computer systems. *ACM Transactions on Information Systems (TOIS)* **14**(3) 330–347.

Goldbach, Carina, Jörn Sickmann, Thomas Pitz. 2022. Sequential decision bias–evidence from grading exams. *Applied Economics* **54**(32) 3727–3739.

Gonzalez-Betancor, Sara M., Alexia J. Lopez-Puig, M. Eugenia Cardenal. 2021. Digital inequality at home. the school as compensatory agent. *Computers & Education* **168** 104195.

Gueguen, Nicolas. 2017. "mr de bussy" is more employable than "mr bussy": the impact of a particle associated with the surname of an applicant in a job application evaluation context. *Names* **65**(2) 104–111.

Hannak, Aniko, Caludia Wagner, David Garcia, Alan Mislove, Markus Strohmaier, Christo Wilson. 2017. Bias in online freelance marketplaces: evidence from taskrabbit and fiverr. *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 1914–1933.

Hill, Andrew J., Weina Zhou. 2023. Peer discrimination in the classroom and academic achievement. *Journal of Human Resources* **58**(3) 0919–10460R3.

Ibanez, Maria R., Michael W. Toffel. 2020. How scheduling can bias quality assessment: evidence from food-safety inspections. *Management Science* **66**(6) 2396–2416.

Iraj, Hamideh, Anthea Fudge, Huda Khan, Margaret Faulkner, Abelardo Pardo, Vitomir Kovanovic. 2021. Narrowing the feedback gap: examing student engagement with personalized and actionable feedback messages. *Journal of Learning Analytics* **8**(3) 101–116.

Jackson, Michelle, Brian Holzman. 2020. A century of educational inequality in the united states. *Proceedings of the National Academy of Sciences* **117**(32) 19108–19115.

Krishna, Aradhna, A. Yesim Orhun. 2022. Gender (still) matters in business school. *Journal of Marketing Research* **59**(1) 191–210.

Lambrecht, Anja, Catherine Tucker. 2019. Algorithmic bias? an empirical study of apparent gender-based discrimination in the display of stem career ads. *Management science* **65**(7) 2966–2981.

Lavy, Victor, Rigissa Megalokonomou. 2019. Persistency in teachers' grading bias and effects on longer-term outcomes: university admissions exams and choice of field of study. *NBER Working Paper 26021* .

Leslie, David. 2020. Understanding bias in facial recognition technologies. *arXiv preprint: 2010.07023* .

Lindsey, Peggy, Deborah J. Crusan. 2011. How faculty attitudes and expectations toward student nationality affect writing assessment. *Across the Disciplines: A Journal of Language, Learning, and Academic Writing* **8**.

Mason, Benjamin A., Adalet Baris Gunersel, Emilie A. Ney. 2014. Cultural and ethnic bias in teacher ratings of behavior: a criterion-focused review. *Osychology in the Schools* **51**(10) 1017–1030.

McCullough, Bruce D, Thomas P McWilliams. 2011. Students with the initial "a" don't get better grades. *Journal of Research in Personality* **45**(3) 340–343.

Mehrabi, Ninareh, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys* **54**(6) 115.

Metaxa-kakavouli, Danae, Kelly Wang, James A. Landay, Jeff Hancock. 2018. Gender-inclusive design: sense of belonging and bias in web interfaces. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 614.

Obermeyer, Ziad, Brian Powers, Christine Vogeli, Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* **366**(6464) 447–453.

Pastushenko, Olena, Cale Passmore. 2023. Play and resistance: intersecting identities and implicit biases in gamified educational tools. *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. 229.

Pong, Suet-Ling. 1997. Family structure, school context, and eighth-grade math and reading achievement. *Journal of Marriage and Family* **59**(3) 734–746.

Protivinsky, Tomas, Daniel Munich. 2018. Gender bias in teachers' grading: what is in the grade. *Studies in Educational Evaluation* **59**(2018) 141–149.

Quarles, Christopher L, Ceren Budak, Paul Resnick. 2020. The shape of educational inequality. *Science Advances* **6**(29) eaaz5954.

Quinn, David M. 2020. Experimental evidence on teachers' racial bias in student evaluation: the role of grading scales. *Educational Evaluation and Policy Analysis* **42**(3) 375–392.

Ravenscroft, Susan P., Frank A. Buckless. 1992. The effect of grading policies and student gender on academic performance. *Journal of Accounting Education* **10** 163–179.

Redding, Christopher. 2019. A teacher like me: a review of the effect of student-teacher racial/ethnic matching on teacher perceptions of students and student academic adn behavioral outcomes. *Review of Educational Research* **89**(4) 499–535.

Regan, Priscilla M., Jolene Jesse. 2019. Ethical challenges of edtech, big data and personalized learning: twenty-first century student sorting and tracking. *Ethics and Information Technology* **21** 167–179.

Sabnis, Sunil, Renzhe Yu, René F Kizilcec. 2022. Large-scale student data reveal sociodemographic gaps in procrastination behavior. *Proceedings of the Ninth ACM Conference on Learning@ Scale*. 133–141.

Santos, Jário, Ig Bittencourt, Marcelo Reis, Geiser Chalco, Seiji Isotani. 2022. Two billion registered students affected by stereotyped educational environments: an analysis of gender-based color bias. *Humanities and Social Sciences Communications* **9**(1) 1–16.

Scardamalia, Marlene, Carl Bereiter. 2008. Pedagogical biases in educational technologies. *Educational Technology* **48**(3) 3–11.

Seethal. 2022. Sentiment analysis generic dataset. URL https://huggingface.co/Seethal/sentiment_analysis_generic_dataset.

Shevlin, Mark, Mark N. Davies. 1997. Alphabetical listing and citation rates. *Nature* **388** 14.

Sievertsen, Hans Henrik, Francesca Gino, Marco Pivoesan. 2016. Cognitive fatigue influences students' performance on standarized tests. *Proceedings of the National Academy of Sciences* **113**(10) 2621–2624.

Simonsohn, Uri, Francesca Gino. 2013. Daily horizons: Evidence of narrow bracketing in judgment from 10 years of mba admissions interviews. *Psychological science* **24**(2) 219–224.

Stefanova, Vasilena, Ioana Latu, Laura Taylor. 2023. What is in a name? exploring perceptions of surname change in hiring evaluations in academic. *Social Sciences* **12**(2) 95.

Tregenza, Tom. 1997. Darwin a better name than wallace? *Nature* **385** 480.

Troyna, Barry. 2012. *Racial inequality in education*. Taylor & Francis.

Turnbull, Darren, Ritesh Chugh, Jo Luck. 2020. Learning management systems, an overview. *Encyclopedia of education and information technologies* 1052–1058.

Vives, Marc-Lluís, Tania Fernandez-Navia, Jordi J Teixidó, Miquel Serra-Burriel. 2021. Lenience breeds strictness: The generosity-erosion effect in hiring decisions. *Science Advances* **7**(17) eabe2045.

Wang, Jingyan, Ashwin Pananjady. 2022. Modeling and correcting bias in sequential evaluation. *arXiv preprint arXiv:2205.01607* .

Wang, Zhihan (Helen), Jun Li, Di (Andrew) Wu. 2023. Mind the gap: gender disparity in online learning interactions. *Manufacturing & Service Operations Management* **Forthcoming**.

Wang, Zijian, David Jurgens. 2018. It's going to be okay: Measuring access to support in online communities. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 33–45.

Wisniewski, Benedikt, Klaus Zierer, John Hattie. 2020. The power of feedback revisited: a meta-analysis of educational feedback research. *Frontiers in Psychology* **10**(2019) 3087.

## Sequential Bias and System-Induced Disparity: Online Appendix

# Appendix A: Behavioral Bias Results

**Table A1 Assignment Grades and Grading Order**

| term | Full | Random Grading | No Late Submission | Reverse Grading |
|---|---|---|---|---|
| Grading order 10-20 | -0.0459*** | -0.0577*** | -0.0354*** | -0.0848*** |
| | (0.0022) | (0.0046) | (0.0020) | (0.0055) |
| Grading order 20-30 | -0.0834*** | -0.0952*** | -0.0651*** | -0.1487*** |
| | (0.0032) | (0.0070) | (0.0031) | (0.0083) |
| Grading order 30-40 | -0.1087*** | -0.1228*** | -0.0876*** | -0.1712*** |
| | (0.0042) | (0.0091) | (0.0040) | (0.0119) |
| Grading order 40-50 | -0.1531*** | -0.1768*** | -0.1239*** | -0.2516*** |
| | (0.0062) | (0.0128) | (0.0060) | (0.0189) |
| Grading order 50-60 | -0.2117*** | -0.2616*** | -0.1763*** | -0.3283*** |
| | (0.0104) | (0.0247) | (0.0103) | (0.0330) |
| Black | -0.1200*** | -0.1040*** | -0.1139*** | -0.1364*** |
| | (0.0078) | (0.0144) | (0.0075) | (0.0148) |
| Hispanic | 0.0377*** | 0.0353** | 0.0345*** | 0.0107 |
| | (0.0070) | (0.0126) | (0.0068) | (0.0138) |
| Asian | 0.0255*** | 0.0335** | 0.0264*** | 0.0239* |
| | (0.0056) | (0.0103) | (0.0054) | (0.0107) |
| Other ethnicities | 0.0078* | 0.0099 | 0.0087* | -0.0121 |
| | (0.0035) | (0.0063) | (0.0034) | (0.0074) |
| Initial F-J | 0.0263*** | 0.0181*** | 0.0234*** | 0.0183*** |
| | (0.0021) | (0.0037) | (0.0020) | (0.0047) |
| Initial K-O | 0.0367*** | 0.0171*** | 0.0313*** | 0.0200*** |
| | (0.0022) | (0.0037) | (0.0021) | (0.0045) |
| Initial P-T | 0.0498*** | 0.0240*** | 0.0416*** | 0.0175*** |
| | (0.0024) | (0.0039) | (0.0024) | (0.0048) |
| Initial U-Z | 0.0678*** | 0.0295*** | 0.0592*** | 0.0142* |
| | (0.0030) | (0.0048) | (0.0029) | (0.0058) |
| Native speaker | -0.0138*** | -0.0129*** | -0.0175*** | -0.0151*** |
| | (0.0023) | (0.0039) | (0.0022) | (0.0041) |
| Female | 0.0814*** | 0.0809*** | 0.0785*** | 0.0676*** |
| | (0.0020) | (0.0033) | (0.0019) | (0.0037) |
| International student | 0.0215** | 0.0099 | 0.0229*** | 0.0228 |
| | (0.0072) | (0.0137) | (0.0069) | (0.0144) |
| Domestic Minority | 0.0019 | -0.0097 | 0.0015 | -0.0013 |
| | (0.0051) | (0.0095) | (0.0050) | (0.0104) |
| Domestic under representative minority | -0.1164*** | -0.1054*** | -0.1098*** | -0.0865*** |
| | (0.0064) | (0.0123) | (0.0063) | (0.0126) |
| High school GPA | 0.0206*** | 0.0191*** | 0.0200*** | 0.0214*** |
| | (0.0007) | (0.0012) | (0.0007) | (0.0013) |
| Single Parent Family | -0.0256*** | -0.0219*** | -0.0244*** | -0.0226*** |
| | (0.0022) | (0.0042) | (0.0022) | (0.0047) |
| Current GPA | 0.3487*** | 0.3527*** | 0.3357*** | 0.4241*** |
| | (0.0039) | (0.0067) | (0.0038) | (0.0056) |
| Previous term cumulative GPA | 0.0099*** | 0.0067** | 0.0098*** | 0.0114*** |
| | (0.0011) | (0.0021) | (0.0011) | (0.0019) |
| Number of credits (count towards GPA) | -0.2331*** | -0.2549*** | -0.2268*** | -0.2520*** |
| | (0.0050) | (0.0098) | (0.0049) | (0.0083) |
| Number of credits (Do not count for GPA) | -0.1008*** | -0.0964*** | -0.0968*** | -0.1054*** |
| | (0.0023) | (0.0040) | (0.0022) | (0.0038) |
| Class size | 0.0080*** | 0.0060*** | 0.0061*** | 0.0060*** |
| | (0.0002) | (0.0003) | (0.0002) | (0.0003) |
| _cons | -1.0430*** | -1.0665*** | -1.0539*** | -1.5929*** |
| | (0.1852) | (0.2614) | (0.1941) | (0.0526) |
| R-squared | 0.075 | 0.068 | 0.067 | 0.087 |
| Observations | 4357309 | 794648 | 4071254 | 617480 |
| Grader FE | Yes | Yes | Yes | Yes |
| Student Country FE | Yes | Yes | Yes | Yes |
| Student Academic Level FE | Yes | Yes | Yes | Yes |

*Notes:* *** p<0.01, ** p<0.05, * p<0.1.

**Table A2 Textual metrics and Grading Order**

| term | Sentiment | Politeness | Student Question | Regrade Request |
|---|---|---|---|---|
| Grading order 10-20 | -0.0423*** | -0.0242*** | 0.0008*** | 0.0006*** |
| | (0.0026) | (0.0025) | (0.0001) | (0.0000) |
| Grading order 20-30 | -0.0686*** | -0.0303*** | 0.0015*** | 0.0011*** |
| | (0.0040) | (0.0039) | (0.0001) | (0.0001) |
| Grading order 30-40 | -0.0925*** | -0.0501*** | 0.0020*** | 0.0015*** |
| | (0.0056) | (0.0054) | (0.0001) | (0.0001) |
| Grading order 40-50 | -0.1030*** | -0.0479*** | 0.0030*** | 0.0023*** |
| | (0.0078) | (0.0075) | (0.0002) | (0.0001) |
| Grading order 50-60 | -0.1444*** | -0.0787*** | 0.0034*** | 0.0032*** |
| | (0.0119) | (0.0128) | (0.0003) | (0.0003) |
| Black | -0.0175* | -0.0130 | 0.0008*** | 0.0005** |
| | (0.0080) | (0.0074) | (0.0002) | (0.0002) |
| Hispanic | 0.0067 | 0.0043 | 0.0003 | 0.0002 |
| | (0.0077) | (0.0073) | (0.0002) | (0.0002) |
| Asian | 0.0009 | 0.0011 | 0.0003* | 0.0001 |
| | (0.0058) | (0.0057) | (0.0002) | (0.0001) |
| Other ethnicities | -0.0047 | -0.0022 | 0.0002* | 0.0002** |
| | (0.0040) | (0.0038) | (0.0001) | (0.0001) |
| Initial F-J | 0.0177*** | 0.0113*** | -0.0002* | -0.0001 |
| | (0.0024) | (0.0024) | (0.0001) | (0.0001) |
| Initial K-O | 0.0269*** | 0.0144*** | -0.0004*** | -0.0003*** |
| | (0.0027) | (0.0025) | (0.0001) | (0.0001) |
| Initial P-T | 0.0312*** | 0.0165*** | -0.0007*** | -0.0004*** |
| | (0.0029) | (0.0028) | (0.0001) | (0.0001) |
| Initial U-Z | 0.0448*** | 0.0290*** | -0.0008*** | -0.0005*** |
| | (0.0035) | (0.0034) | (0.0001) | (0.0001) |
| Native speaker | -0.0056** | -0.0103*** | -0.0001 | -0.0000 |
| | (0.0022) | (0.0022) | (0.0000) | (0.0000) |
| Female | 0.0326*** | 0.0323*** | 0.0001 | -0.0001*** |
| | (0.0019) | (0.0019) | (0.0000) | (0.0000) |
| International student | 0.0044 | -0.0055 | 0.0003 | -0.0000 |
| | (0.0078) | (0.0074) | (0.0002) | (0.0002) |
| Domestic Minority | 0.0021 | -0.0113* | -0.0002 | -0.0002 |
| | (0.0056) | (0.0054) | (0.0001) | (0.0001) |
| Domestic under representative minority | -0.0179* | -0.0055 | 0.0000 | -0.0001 |
| | (0.0070) | (0.0066) | (0.0002) | (0.0001) |
| High school GPA | 0.0041*** | 0.0028*** | -0.0001*** | -0.0001*** |
| | (0.0007) | (0.0006) | (0.0000) | (0.0000) |
| Single Parent Family | -0.0031 | -0.0018 | 0.0001 | 0.0001 |
| | (0.0025) | (0.0025) | (0.0001) | (0.0000) |
| Current GPA | 0.0520*** | 0.0416*** | 0.0004*** | 0.0003*** |
| | (0.0021) | (0.0020) | (0.0000) | (0.0000) |
| Previous term cumulative GPA | 0.0026** | 0.0013 | -0.0001** | -0.0000 |
| | (0.0009) | (0.0009) | (0.0000) | (0.0000) |
| Number of credits (count towards GPA) | -0.0469*** | -0.0410*** | -0.0002** | -0.0001 |
| | (0.0036) | (0.0032) | (0.0001) | (0.0001) |
| Number of credits (Do not count for GPA) | -0.0175*** | -0.0121*** | -0.0000 | -0.0000 |
| | (0.0016) | (0.0015) | (0.0000) | (0.0000) |
| Class size | 0.0028*** | 0.0017*** | -0.0000*** | -0.0000*** |
| | (0.0002) | (0.0002) | (0.0000) | (0.0000) |
| _cons | -1.7233*** | -1.0996*** | 0.0035* | 0.0004 |
| | (0.4850) | (0.1873) | (0.0015) | (0.0022) |
| R-squared | 0.059 | 0.045 | -0.001 | -0.001 |
| Observations | 1583018 | 1852079 | 4369627 | 4369627 |
| Grader FE | Yes | Yes | Yes | Yes |
| Student Country FE | Yes | Yes | Yes | Yes |
| Student Academic Level FE | Yes | Yes | Yes | Yes |

*Notes:* *** p<0.01, ** p<0.05, * p<0.1.

## Appendix B: System Design Results

**Table B1 Grading Mode and Surname Initial Grade Disparity**

| term | Online VS. Random | Group Assignments |
|---|---|---|
| Initial F-J | 0.028593*** | -0.015044** |
| | (0.004908) | (0.007348) |
| Initial K-O | 0.028146*** | -0.010043 |
| | (0.004844) | (0.007056) |
| Initial P-T | 0.032223*** | -0.004308 |
| | (0.005109) | (0.007471) |
| Initial U-Z | 0.045618*** | -0.005702 |
| | (0.006297) | (0.008807) |
| initial_cate_0_seq_grade | 0.016122*** | -0.008184 |
| | (0.003948) | (0.018368) |
| initial_cate_1_seq_grade | -0.006394 | -0.004395 |
| | (0.004308) | (0.017365) |
| initial_cate_2_seq_grade | -0.010892*** | -0.015748 |
| | (0.004188) | (0.018167) |
| initial_cate_3_seq_grade | -0.015057*** | -0.032171* |
| | (0.004642) | (0.018764) |
| initial_cate_4_seq_grade | -0.016736*** | -0.001963 |
| | (0.005995) | (0.026641) |
| Black | -0.127282*** | -0.052303** |
| | (0.009938) | (0.025208) |
| Hispanic | 0.036359*** | 0.022334 |
| | (0.008837) | (0.023313) |
| Asian | 0.028108*** | 0.044626** |
| | (0.006964) | (0.017887) |
| Other ethnicities | 0.009214** | 0.012228 |
| | (0.004428) | (0.010853) |
| Native speaker | -0.013899*** | -0.013648* |
| | (0.002744) | (0.007522) |
| Female | 0.079947*** | 0.065325*** |
| | (0.002327) | (0.005840) |
| International student | 0.021632** | -0.001355 |
| | (0.008782) | (0.020756) |
| Domestic Minority | 0.002521 | -0.040455** |
| | (0.006461) | (0.016820) |
| Domestic under representative minority | -0.114571*** | -0.029904 |
| | (0.008088) | (0.022656) |
| High school GPA | 0.020128*** | 0.008333*** |
| | (0.000862) | (0.002142) |
| Single Parent Family | -0.025723*** | -0.005855 |
| | (0.002811) | (0.008686) |
| Current GPA | 0.363669*** | 0.201122*** |
| | (0.004323) | (0.007875) |
| Previous term cumulative GPA | 0.012453*** | -0.003151 |
| | (0.001317) | (0.003247) |
| Number of credits (count towards GPA) | -0.239224*** | -0.161582*** |
| | (0.005801) | (0.012701) |
| Number of credits (Do not count for GPA) | -0.108103*** | -0.040986*** |
| | (0.002633) | (0.005131) |
| Class size | 0.006403*** | 0.003314*** |
| | (0.000164) | (0.000491) |
| _cons | -1.336005*** | -0.540978 |
| | (0.211101) | (0.361476) |
| R-squared | 0.077 | 0.031 |
| Observations | 2201181 | 222515 |
| Grader FE | Yes | Yes |
| Student Country FE | Yes | Yes |
| Student Academic Level FE | Yes | Yes |

*Notes:* *** p<0.01, ** p<0.05, * p<0.1.

**Table B2 Grading Mode and Surname Initial Grade Disparity: Text Metrics**

| term | Sentiment | Politeness | Student Question | Regrade Request |
|---|---|---|---|---|
| Initial F-J | 0.004826 | 0.013184* | -0.000093 | -0.000190 |
| | (0.007495) | (0.006998) | (0.000185) | (0.000154) |
| Initial K-O | 0.016887** | 0.011175 | -0.000004 | -0.000258* |
| | (0.007396) | (0.007129) | (0.000189) | (0.000153) |
| Initial P-T | 0.014715* | 0.007326 | -0.000369* | -0.000271* |
| | (0.007531) | (0.007396) | (0.000189) | (0.000158) |
| Initial U-Z | 0.027781*** | 0.029008*** | -0.000165 | -0.000043 |
| | (0.009303) | (0.008956) | (0.000245) | (0.000207) |
| initial_cate_0_seq_grade | -0.003952 | -0.004023 | -0.000311** | -0.000291*** |
| | (0.006129) | (0.005974) | (0.000124) | (0.000106) |
| initial_cate_1_seq_grade | -0.007305 | -0.013102** | -0.000026 | 0.000093 |
| | (0.006697) | (0.006351) | (0.000141) | (0.000118) |
| initial_cate_2_seq_grade | -0.024685*** | -0.014947** | 0.000053 | 0.000138 |
| | (0.006580) | (0.006505) | (0.000136) | (0.000107) |
| initial_cate_3_seq_grade | -0.028547*** | -0.012368* | 0.000319** | 0.000208* |
| | (0.007094) | (0.006936) | (0.000143) | (0.000119) |
| initial_cate_4_seq_grade | -0.031640*** | -0.023037** | 0.000209 | 0.000112 |
| | (0.009141) | (0.009100) | (0.000214) | (0.000177) |
| Black | -0.015116 | -0.005945 | 0.001085*** | 0.000538** |
| | (0.011342) | (0.010441) | (0.000259) | (0.000217) |
| Hispanic | 0.000004 | 0.000196 | 0.000300 | 0.000127 |
| | (0.010630) | (0.009971) | (0.000233) | (0.000195) |
| Asian | -0.003483 | 0.000843 | 0.000264 | 0.000135 |
| | (0.008293) | (0.008029) | (0.000190) | (0.000156) |
| Other ethnicities | -0.004561 | -0.000901 | 0.000192 | 0.000231** |
| | (0.005776) | (0.005445) | (0.000124) | (0.000102) |
| Native speaker | -0.006185** | -0.012628*** | -0.000114* | -0.000071 |
| | (0.003028) | (0.002982) | (0.000060) | (0.000049) |
| Female | 0.033886*** | 0.032074*** | 0.000050 | -0.000150*** |
| | (0.002698) | (0.002662) | (0.000056) | (0.000045) |
| International student | 0.011225 | -0.003229 | 0.000415 | -0.000056 |
| | (0.011073) | (0.010158) | (0.000286) | (0.000239) |
| Domestic Minority | 0.009136 | -0.007737 | -0.000170 | -0.000143 |
| | (0.008031) | (0.007720) | (0.000178) | (0.000147) |
| Domestic under representative minority | -0.019197** | -0.010745 | -0.000161 | -0.000105 |
| | (0.009636) | (0.009098) | (0.000215) | (0.000187) |
| High school GPA | 0.003232*** | 0.003090*** | -0.000043** | -0.000039** |
| | (0.000930) | (0.000865) | (0.000019) | (0.000016) |
| Single Parent Family | -0.000668 | -0.006927** | 0.000046 | 0.000074 |
| | (0.003549) | (0.003437) | (0.000074) | (0.000061) |
| Current GPA | 0.054735*** | 0.043725*** | 0.000309*** | 0.000301*** |
| | (0.002893) | (0.002630) | (0.000052) | (0.000043) |
| Previous term cumulative GPA | 0.002846** | 0.002007* | -0.000028 | -0.000014 |
| | (0.001209) | (0.001115) | (0.000025) | (0.000020) |
| Number of credits (count towards GPA) | -0.048448*** | -0.039720*** | -0.000012 | -0.000141** |
| | (0.004573) | (0.004150) | (0.000086) | (0.000070) |
| Number of credits (Do not count for GPA) | -0.020036*** | -0.012645*** | 0.000024 | 0.000011 |
| | (0.002287) | (0.002056) | (0.000045) | (0.000037) |
| Class size | 0.001458*** | 0.001017*** | -0.000005** | -0.000006*** |
| | (0.000218) | (0.000224) | (0.000002) | (0.000002) |
| _cons | -1.950129*** | -0.976012*** | 0.004254*** | 0.000423 |
| | (0.571460) | (0.301577) | (0.001518) | (0.002451) |
| R-squared | 0.059 | 0.045 | -0.003 | -0.003 |
| Observations | 736423 | 873748 | 2209354 | 2209354 |
| Grader FE | Yes | Yes | Yes | Yes |
| Student Country FE | Yes | Yes | Yes | Yes |
| Student Academic Level FE | Yes | Yes | Yes | Yes |

*Notes:* *** p<0.01, ** p<0.05, * p<0.1.