What is Best for Students, Numerical Scores or Letter Grades?

Evi Micha
Department of Computer Science
University of Toronto
emicha@cs.toronto.edu

Shreyas Sekar
Department of Management
University of Toronto
shreyas.sekar@rotman.toronto.edu

Nisarg Shah
Department of Computer Science
University of Toronto
nisarg@cs.toronto.edu

Abstract

We study letter grading schemes, which are routinely employed for evaluating student performance. Typically, a numerical score obtained through one or more evaluations is converted into a letter grade (e.g., A+, B-, etc.) by associating a disjoint interval of numerical scores to each letter grade.

We propose the first model for understanding the (de)motivational effects of such grading on the students and, consequently, on their performance in future evaluations. We use the model to compare different letter grading schemes to numerical scoring, in which the score is not converted to any letter grade (equivalently, every score is its own letter grade).

Theoretically, we identify realistic conditions under which numerical scoring is better than any uniform letter grading scheme. Our experiments confirm that this holds under even weaker conditions, but also identify other realistic conditions under which uniform letter grading schemes outperform numerical scoring.

1 Introduction

Student evaluations and grading play an integral and influential role in every individual’s academic experience. Naturally, there has been widespread debate among researchers and policy-makers about the efficacy of various grading systems such as letter v.s. number grades. For instance, coarse-grained grading schemes (i.e., letter grades) are considered to be less noisy indicators of performance, and stronger signals of status, and consequently, are the norm in North American universities. At the same time, there is also growing awareness that the grade itself affects performance independent of student ability, i.e., the grades are “not just an output of the educational process, they may also be an input” Gray and Bunte [2022]. For example, empirical evidence suggests the disclosure of midterm grades may motivate or demotivate students to perform better in a future exam, controlling for other effects. In light of this evidence, it is clear that the design of a grading system must be a deliberate choice that takes into account student welfare in addition to other extraneous factors Guskey [2011].

In this work, we take an analytical approach and study the design of an optimal grading system with a particular focus on numeric v.s. uniform letter grades¹ schemes. As far as we are aware, this is among the first works that looks at the problem of designing a grading scheme with the explicit objective of

¹We use the term uniform letter grades to refer to letter grading schemes where each letter grade corresponds to an equal sized score range, e.g., [90, 100] → A+, [80, 90] → A-, and so on.
improving student performance in future tests. Our model captures the impact of grades on future performance via two well-motivated effects:

1. **Anchoring**: In any given test, students anchor themselves to a specific score or performance level based on their intrinsic ability which directly impacts their performance.

2. **(De)Motivation**: When the student’s actual score falls above (below) their anchor, they get (de)motivated and subsequently, their expectation increases (decreases) for future tests.

In this regard, our work departs from other papers in this area, where students are often modelled as status-maximizers Dubey and Geanakoplos [2010], i.e., their intrinsic motivation for a better grade stems from a desire to rank above their fellow students. Our model does not induce any artificial scarcity (status) and instead the fundamental friction is result of noisy performance and how the same grading rule affects different students differently.

To better illustrate how different grading schemes impact student performance under our model, consider the case of two students with the same intrinsic ability (score anchors) \( q_1 = q_2 = 85 \). Due to random factors, the first student’s score in the midterm is given by \( s_1 = 81 \) while the second student matches expectations and scores \( s_2 = 85 \). In this case, disclosing the numeric score may demotivate student 1, leading to an anchor for the final exam that is lower than 85. On the other hand, under a coarser scheme, both students could receive a letter grade (say) \( A- \) capturing all scores in the range \([80, 90]\), which limits the adverse effect on future performance. At the same time, a third student who’s intrinsic ability is \( q_3 = 91 \) and whose midterm score is \( s_3 = 89 \) may also be bracketed into the same letter grade \( A- \), leading to severe demotivation. In this scenario, the disclosure of the numeric grade informs the student that their score was actually close to their anchor.

Building on the ideas presented in this example, we develop a framework to compare various grading systems in any environment with multiple sequential tests. This includes evaluations within a course, e.g., a midterm followed by a final exam but also grading across related courses, i.e., a student taking Calculus 101 followed by Calculus 102. In all of these scenarios, we prove prescriptive insights on when it is optimal to use numeric grades as opposed to uniform letter grades. Specifically, we show that when students’ scores are highly likely to be close to their anchors, numeric grading always outperforms any arbitrary choice of uniform letter grades.

**Our results.** In this work, we compare different grading schemes in terms of how a student’s performance in a subsequent evaluation may be affected by her grade on an earlier evaluation. A student is demotivated if her grade is less than her true quality and is motivated, otherwise. Our goal is to maximize student’s grade in a subsequent evaluation. We consider the numerical scoring scheme, where the student learns her exact score in an evaluation, to uniform letter grading schemes where the interval of grades is partitioned into \( T \) equal length-intervals and each interval is represented by its midpoint, i.e. the student learns the midpoint of the interval that her score falls into.

In Section 3, we consider the case that there are two sequential evaluations and the true quality of the students is drawn from a uniform distribution. Our main result is that when the demotivational effect is larger then the motivational effect, then numerical scoring is better than any uniform letter grading, while exactly the opposite is true when the demotivational effect is smaller than the motivational effect.

**Related work.** There are a rich literature that compares different grading schemes over different objectives. However, all these objectives are very different with our goal in this work.

It has been studied both theoretically and empirically how the effort that the students exert for an evaluation depends on the grading scheme that is used Paredes [2017], BRO [2018], Main and Ost [2014], Czibor et al. [2020]. This comparison is quite intuitive as for example when the grading scheme is based on a pass/fail approach, a student can try as hard as to ensure that pass in an evaluation. Our goal is completely different, as we study how students motivation or demotivation may be affected due to the fact that the grading scheme may not be quite descriptive. Another related work is Sikora [2022] which also compares different grading schemes but the goal there is given scores that the students have obtain through the execution of a course to assign final grades to the students using a grading scheme that first convey as much information as possible for students performances and second is not affected by random factors that may affect students scores. It has been also studied how different grading schemes may impact students psychological well-being and stress Rohe et al. [2006], Bloodgood et al. [2020].
2 Model

Define \( |k| = \{1, \ldots, k\} \) for \( k \in \mathbb{N} \). We introduce a model in which the grading scheme used in one evaluation can motivate or demotivate students, affecting their performance in future evaluations.

**True qualities.** A student begins with an intrinsic (true) quality \( q \) drawn from a (nonatomic) prior \( \mathcal{Q} \) with probability density function (PDF) \( f_\mathcal{Q}(\cdot) \). For simplicity, let the support of \( \mathcal{Q} \) be \([0, 1]\).

**Scores.** There is a score model \( \mathcal{S} \) such that the numerical performance (score) of a student with true quality \( q \) in the first evaluation, denoted \( s \in [0, 1] \), is drawn from the (nonatomic) distribution \( \mathcal{S}(q) \) with PDF \( f_\mathcal{S}(\cdot; q) \). We focus on score models in which the expected score of a student is equal to their true quality, i.e., \( \mathbb{E}_{s \sim \mathcal{S}(q)}[s] = q \) for all \( q \in [0, 1] \).

**Grades.** A grading scheme is a function \( B : [0, 1] \rightarrow [0, 1] \) that maps the score to a grade.

**Letter grading.** A letter grading scheme \( \mathcal{B}_\mathcal{L} \) is specified by a vector \( \mathbf{c} = (c_0 = 0, c_1, \ldots, c_{T-1}, c_T = 1) \), for some \( T \in \mathbb{N} \) (referred to as the number of grades) and \( c_i \geq c_{i-1} \) for all \( i \in [T] \), and is given by \( B_\mathcal{L}(s) = \frac{c_i - s}{c_i - c_{i-1}} \) for all \( i \in [T] \) and \( s \in [c_{i-1}, c_i) \). That is, it partitions \([0, 1]\) into finitely many disjoint intervals (one for each grade) and maps a score to the midpoint of the interval containing it.

**Uniform letter grading.** We are particularly interested in the uniform letter grading (ULG) scheme. For a given number of grades \( T \in \mathbb{N} \), uniform letter grading with \( T \) grades, denoted \( \text{ULG}_{T} \), is specified by \( c_i = i/T \) for each \( i \in [T] \). In other words, it partitions \([0, 1]\) into \( T \) equal-length intervals. We will use \( \Delta(T) = 1/T \) to denote the length of the interval, dropping \( T \) from the argument when it is clear from the context. Formally, we have that for all \( s \in [0, 1] \),

\[
\text{ULG}_T(s) = \left(\lfloor s/\Delta \rfloor + 1/2 \right) \cdot \Delta.
\]

For instance, \( \text{ULG}_{10} \) maps all scores in \([0, 0.1)\) to \(0.05\), all scores in \([0.1, 0.2)\) to \(0.15\), and so on.

**Numerical scoring.** We will compare (uniform) letter grading to numerical scoring (NS), given by \( \text{NS}(s) = s \) for all \( s \in [0, 1] \). Under numerical scoring, scores are not rounded to any grades. This can also be viewed as the limit of uniform letter grading with \( T \rightarrow \infty \) grades.

**(De)motivation.** The grades affect students’ level of motivation in subsequent evaluations. Under grading scheme \( B \), a student compares their true quality \( q \) to the obtained grade \( B(s) \). If the grade is higher than the true quality, the student experiences a motivational boost, but in the converse case, gets demotivated. We model this by assuming that the effective true quality of the student for the next evaluation changes to \( q' = q + h(q, B(s)) \), where

\[
h(q, B(s)) = \begin{cases} 
\alpha_m \cdot (B(s) - q), & \text{if } B(s) \geq q, \\
-\alpha_d \cdot (q - B(s)), & \text{if } B(s) < q.
\end{cases}
\]

We refer to \( \alpha_m, \alpha_d \in \mathbb{R}_{\geq 0} \) as motivation and demotivation coefficients, respectively. Note that the amount of (de)motivation is proportional to the difference between the obtained grade and the true quality. In the next evaluation, the student then obtains a score \( s' \) drawn from \( \mathcal{S}(q') \). We remark that when \( \alpha_m, \alpha_d \in [0, 1] \), we automatically have \( q' \in [0, 1] \); thus, we focus on this range of parameters.\(^3\)

**Goal.** Intuitively, we are interested in choosing grading schemes that maximize the average performance of the students. Note that

\[
\mathbb{E}[s + s'] = \mathbb{E}[s] + \mathbb{E}[s'] = \mathbb{E}[q] + \mathbb{E}[q'] = 2 \cdot \mathbb{E}[q] + \mathbb{E}[h(q, B(s))],
\]

where the first transition is due to linearity of expectation and the second transition is because the expected score is always equal to the true quality in our model. Hence, two natural objectives — maximizing the average score in the final evaluation \( \mathbb{E}[s'] \) and maximizing the average score across both evaluations \( \mathbb{E}[(s + s')/2] \) — both reduce to the same objective of maximizing \( \text{perf}(B) \overset{\triangleq}{=} \mathbb{E}_{q \sim \mathcal{Q}, s \sim \mathcal{S}(q)}[h(q, B(s))] \) because \( \mathbb{E}[q] \), the mean of the true quality prior \( \mathcal{Q} \), is independent of the grading scheme used. Hence, we focus on analyzing \( \text{perf}(B) \) and comparing different grading schemes by this objective.

Later, in Section 4, we study extensions of this model where students go through more than two evaluations and analyze other objective functions such as the probability that a student with lower true quality ends up obtaining a higher grade, which can be viewed as a measure of unfairness.\(^4\)

---

\(^3\)Because we assume nonatomic distributions, it does not matter what \( \text{ULG}_T(1) \) is. We will use the convention that \( \text{ULG}_T(1) = 1 \).

\(^4\)In principle, one can also use larger coefficients and truncate \( q' \) to lie in \([0, 1]\).
3 Uniform Letter Grading vs Numerical Scoring

In this section, we derive theoretical results that compare the average student performance under uniform letter grading schemes to that under numerical scoring. We identify conditions under which numerical scoring is better than any uniform letter grading scheme, and conditions under which the converse holds. Let us begin by introducing two useful definitions.

**Definition 1** (Jointly Symmetric Distributions). We say that the true quality prior $Q$ and the score model $S$ are jointly symmetric if $f_Q(q) \cdot f_S(s; q) = f_Q(1-q) \cdot f_S(1-s; 1-q)$ for all $s, q \in [0, 1]$.

Joint symmetry requires that true qualities and scores are symmetric across $[0, 1]$. That is, the probability of having true quality $q$ and receiving score $s$ should be the same as the probability of having true quality $1-q$ and receiving score $1-s$. If the true quality prior is uniform, then this means the score distribution $S(q)$ should be the mirror image of the score distribution $S(1-q)$. Note that joint symmetry does not necessarily require symmetry of the “noise” contained in the score compared to the true quality. For example, we do not necessarily need $f_S(s = 0.4; q = 0.5) = f_S(s = 0.6; q = 0.5)$.

**Definition 2** (Symmetric Grading Scheme). We say that a grading scheme $B$ is symmetric if $B(1-s) = 1-B(s)$ for all $s \in [0, 1]$.

The reader can check that numerical scoring (NS) and uniform letter grading schemes (ULG$_T$ for any $T \in \mathbb{N}$) are symmetric.

Our first result shows that under such symmetry, the performance of the grading scheme is linear in the difference between the motivation and demotivation coefficients. As we later show in Corollary 1, this allows us to compare numerical scoring to uniform letter grading.

**Theorem 1.** When the true quality prior $Q$ and the score model $S$ are jointly symmetric, and the grading scheme $B$ is symmetric, then we have

$$\text{perf}(B) = \frac{\alpha_m - \alpha_d}{2} \cdot \mathbb{E}_{q \sim Q, s \sim S(q)} \left[ |q - B(s)| \right].$$

**Proof.** Note that due to $Q$ and $S$ being jointly symmetric, the pairs $(q, s)$ and $(1-q, 1-s)$ are sampled with equal density. Hence, we have that

$$\mathbb{E} \left[ h(q, B(s)) \right] = \frac{1}{2} \cdot \mathbb{E} \left[ h(q, B(s)) + h(1-q, B(1-s)) \right].$$

Due to the symmetry of the grading scheme, we have $B(1-s) = 1-B(s)$, which implies that the two terms $h(q, B(s))$ and $h(1-q, B(1-s))$ are motivation and demotivation by the same amount (but with different coefficients). Hence,

$$\mathbb{E} \left[ h(q, B(s)) + h(1-q, B(1-s)) \right] = (\alpha_m - \alpha_d) \cdot \mathbb{E}_{q \sim Q, s \sim S(q)} \left[ |q - s| \right].$$

Plugging this into Equation (2), we obtain the desired result.

**Corollary 1.** Assume that the true quality prior $Q$ and the score model $S$ are jointly symmetric. Then, all symmetric grading schemes have equal performance when $\alpha_m = \alpha_d$. Further, when $\alpha_m \neq \alpha_d$, for every $T \in \mathbb{N}$ one of the following conditions holds.

1. Uniform letter grading with $T$ grades is at least as good as numerical scoring when $\alpha_m > \alpha_d$, and the converse holds when $\alpha_m < \alpha_d$.

2. Uniform letter grading with $T$ grades is at least as good as numerical scoring when $\alpha_m < \alpha_d$, and the converse holds when $\alpha_m > \alpha_d$.

**Proof.** The first claim regarding $\alpha_m = \alpha_d$ follows immediately from Equation (1) in Theorem 1. For the second claim regarding $\alpha_m \neq \alpha_d$, note that the comparison between numerical scoring and uniform letter grading with $T$ buckets reduces to the sign of $\mathbb{E}[|q - \text{NS}(s)| - |q - \text{ULG}_T(s)|]$, and depending on this sign, one of the two statements in the corollary holds.
Corollary 1 tells us that having equal motivation and demotivation coefficients ($\alpha_m = \alpha_d$) is the turning point: between uniform letter grading with a fixed number of grades and numerical scoring, one is better when $\alpha_m < \alpha_d$ but the other becomes better when $\alpha_m > \alpha_d$. But it does not tell us which one is better in each case.

Our next result identifies a sufficient condition under which this dilemma is settled: uniform letter grading is better when $\alpha_m > \alpha_d$ and numerical scoring is better when $\alpha_m < \alpha_d$. To introduce this sufficient condition, we need to define the following natural property of the score model.

**Definition 3 (Single-Peaked Score Model).** We say that the score model $S$ is single-peaked if, for every $q \in [0, 1]$, $f_S(q; q)$ is single-peaked with the peak at $q$, i.e., $f(s; q) \leq f(s'; q)$ for all $s \leq s' \leq q$ and $s \geq s' \geq q$.

Intuitively, in a single-peaked score model, scores closer to the true quality are more likely than scores farther from the true quality.

**Theorem 2.** Fix any $T \in \mathbb{N}$. Define $D = \{(q, s) : \text{ULG}_T(q) = \text{ULG}_T(s)\}$. Assume that the true quality prior $Q$ and the score model $S$ satisfy the following.

- $Q$ and $S$ are jointly symmetric;
- $S$ is single-peaked; and
- $E\left[|q - s| \mid (q, s) \in D\right] \leq E\left[|q - \text{ULG}_T(s)| \mid (q, s) \in D\right]$.

Then, the first implication of Corollary 1 holds. That is, uniform letter grading with $T$ grades is at least as good as numerical scoring when $\alpha_m > \alpha_d$, the converse holds when $\alpha_m < \alpha_d$, and the two have equal performance when $\alpha_m = \alpha_d$.

Before diving into the proof, let us make a remark regarding the third technical condition in Theorem 2. Note that $D$ is the set of all pairs of true qualities and scores that belong to the same letter grade interval. The technical condition states that, averaged over all such pairs, the true quality is closer to the score than to the midpoint of the interval that they both belong to. Later, we show that this condition is satisfied in two natural cases. Intuitively, if the score distribution is sufficiently concentrated near the true quality, the expected distance between the score and the true quality will be sufficiently small, satisfying the condition. Let us now turn to the proof of Theorem 2.

**Proof.** Given Theorem 1, we simply need to show that $E\left[|q - s| \mid (q, s) \in D\right] \leq E\left[|q - \text{ULG}_T(s)| \mid (q, s) \in D\right]$. We already assume that this holds conditioned on $(q, s) \in D$. Hence, we only need to show that it also holds conditioned on $(q, s) \notin D$. We show that this holds given the additional single-peakedness property.

We show that, conditioned on $(q, s) \notin D$, the desired equation actually holds for every $q \in [0, 1]$, and, therefore, in expectation over $q \sim Q$ as well. Fix arbitrary $q \in [0, 1]$. Note that

$$E\left[|q - s| \mid (q, s) \notin D\right] = \Pr\left[\text{ULG}_T(s) < \text{ULG}_T(q) \mid (q, s) \notin D\right] \cdot E\left[q - s \mid \text{ULG}_T(s) < \text{ULG}_T(q)\right] + \Pr\left[\text{ULG}_T(s) > \text{ULG}_T(q) \mid (q, s) \notin D\right] \cdot E\left[s - q \mid \text{ULG}_T(s) > \text{ULG}_T(q)\right] \leq \Pr\left[\text{ULG}_T(s) < \text{ULG}_T(q) \mid (q, s) \notin D\right] \cdot E\left[q - \text{ULG}_T(s) \mid \text{ULG}_T(s) < \text{ULG}_T(q)\right] + \Pr\left[\text{ULG}_T(s) > \text{ULG}_T(q) \mid (q, s) \notin D\right] \cdot E\left[\text{ULG}_T(s) - q \mid \text{ULG}_T(s) > \text{ULG}_T(q)\right] = E\left[q - \text{ULG}_T(s) \mid (q, s) \notin D\right];$$

where the first transition holds because $[0, 1]^2 \setminus D = \{(q, s) : \text{ULG}_T(s) < \text{ULG}_T(q)\} \cup \{(q, s) : \text{ULG}_T(s) > \text{ULG}_T(q)\}$; and the second transition holds due to linearity of expectation and because the single-peakedness assumption implies

$$E\left[s \mid \text{ULG}_T(s) < \text{ULG}_T(q)\right] \geq E\left[\text{ULG}_T(s) \mid \text{ULG}_T(s) < \text{ULG}_T(q)\right].$$
Then, the first implication of Corollary 1 holds. That is, uniform letter grading with $T$ grades is at least as good as numerical scoring when $\alpha_m > \alpha_d$, the converse holds when $\alpha_m < \alpha_d$, and the two have equal performance when $\alpha_m = \alpha_d$.

The proofs of Theorems 3 and 4 are our most intricate proofs. However, due to space constraints, we have deferred them to the appendix.

Let us however understand the strength of the assumptions in Theorem 3. A natural choice of $S$ under which Assumptions 2 and 3 in Theorem 3 are satisfied is when $\mathcal{S}(q)$ is a symmetric distribution around $q$, i.e., the noise in the score follows a symmetric zero-mean distribution. Further, for such a score model, we have $\gamma = 1$, so Assumption 4 becomes $\Pr[(q, s) \in \mathcal{D}_{\text{same}}] \geq 4 \cdot \Pr[(q, s) \in \mathcal{D}_{\text{opp}}]$. Note that $\mathcal{D}_{\text{same}}$ is the subset of $\mathcal{D}$ in which both $q$ and $s$ are on the same side of the midpoint of the common interval containing them, while in $\mathcal{D}_{\text{opp}}$, they are on the opposite sides of the midpoint. Hence, when the variance of the score distribution is sufficiently small, we can expect $\Pr[(q, s) \in \mathcal{D}_{\text{same}}]$ to be much higher than $\Pr[(q, s) \in \mathcal{D}_{\text{opp}}]$. For further intuition regarding the comparison between $\Pr[(q, s) \in \mathcal{D}_{\text{same}}]$ and $\Pr[(q, s) \in \mathcal{D}_{\text{opp}}]$, see Appendix A in Appendix A.

We noted that under a strong form of symmetry, where the noise distribution is symmetric and zero-mean, Assumptions 2 and 3 of Theorem 3 are met and the constant in Assumption 4 becomes 4. Using very different techniques, we can in fact show that under such strong symmetry, even a constant of 3 would suffice to obtain the same result. This broadens the scope of the result to include slightly less concentrated score distributions. Due to space constraints, we defer the intricate proof of this result to the appendix.
Theorem 4. Fix arbitrary $T \in \mathbb{N}$. Let $D$, $D^{\text{same}}$, and $D^{\text{opp}}$ be defined as in Theorem 3. Assume the following regarding the true quality prior $Q$ and the score model $S$.

1. $Q$ is uniform over $[0, 1]$.
2. (Strong symmetry of $S$) $f_S(s; q) = f_S(s', q')$ whenever $|s - q| = |s' - q'|$.
3. $\Pr[(q, s) \in D^{\text{same}}] \geq 3 \cdot \Pr[(q, s) \in D^{\text{opp}}]$.

Then, the first implication of Corollary 1 holds. That is, uniform letter grading with $T$ grades is at least as good as numerical scoring when $\alpha_m > \alpha_d$, the converse holds when $\alpha_m < \alpha_d$, and the two have equal performance when $\alpha_m = \alpha_d$.

4 Experiments

In this section, we empirically compare different grading schemes. First, we extend our model to allow more than two evaluations. Consider a student with initial true quality $q_1 = q$ who participates in $r$ sequential evaluations. For $j \in [r]$, let $q_j$ and $s_j$ denote her effective true quality and score in evaluation $j$, respectively. Then, $s_j \sim S(q_j)$ for each $j \in [r]$, and for $j \in \{2, \ldots, r\}$, we have:

$$q_j = \begin{cases} q_{j-1} + \alpha_m \cdot (B(s_{j-1}) - q_{j-1}), & \text{if } B(s_{j-1}) \geq q_{j-1}, \\ q_{j-1} - \alpha_d \cdot (q_{j-1} - B(s_{j-1})), & \text{if } B(s_{j-1}) < q_{j-1}. \end{cases}$$

Metrics: For each grading scheme, we measure three metrics:

1. **Final Score Discrepancy**: The expected deviation of the score in the final evaluation from the initial true quality, i.e., $\mathbb{E}[s_r - q_1]$. 
The first two metrics measure the improvement in student performance either by the end or on average the number of evaluations increases. For example, notice that ULG smaller inversion probability than every uniform letter grading. We also see that as the motivation coefficient (right). Interestingly, in both cases, we see that numerical scoring has a number of evaluations (top) and the motivation coefficient (bottom). With αm ∈ {0, 0.1, ..., 0.9, 1} (defaults are 0.2, 0.5, and 0.8). We use r ∈ {2, ..., 25} sequential evaluations (defaults are 2 and 4). When we vary one parameter, we fix the default values of the other parameters to isolate the dependence on the varied parameter.

Results. Figure 1 shows how the expected final score discrepancy changes with the number of evaluations (top) and the motivation coefficient (bottom). With αm = 0.2 < αd, we see that numerical scoring is better than uniform letter grading for a small number of evaluations, matching our theoretical result. But surprisingly, for even a moderate number of evaluations, the comparison quickly flips. The same holds for αm = 0.8 > αd, where numerical scoring starts off being worse than uniform letter grading, but quickly becomes better as the number of evaluations increases. For r = 4 evaluations, there is a sweet spot where numerical scoring is better than uniform letter grading for almost the entire range of αm ∈ [0, 1]. Given this, it is clear that the right choice of a good grading scheme depends not only on the motivation and demotivation coefficients, but also on the number of evaluations.

Figure 2 shows how the inversion probability changes with the number of evaluations (left) and the motivation coefficient (right). Interestingly, in both cases, we see that numerical scoring has a smaller inversion probability than every uniform letter grading. We also see that as the motivation coefficient increases, the probability of inversion also increases. Moreover, the inversion probability of uniform letter grading with a smaller number of grades T seems to increase significantly faster as the number of evaluations increases. For example, notice that ULG4, ULG8 and ULG12 violate the
individual fairness requirement for 25%, 15% and 10% of pairs of the students, respectively, even after 5 consecutive tests. This is quite discouraging as uniform letter grading is often used in practice with 12 or fewer grades. This is a strong indication that we may need to rethink what grading schemes we use in order to ensure that students are not discouraged.

Due to the space constraints, we have presented only a sampling of the most striking empirical observations here, and defer the rest to Appendix C. An interesting observation there is that average score discrepancy turns out to be a very different metric than final score discrepancy.

5 Discussion

Our work takes the first step towards proposing a statistical model of the psychological impact of letter grading schemes on student performance and using it to compare letter grading schemes to numerical scoring. We view our work as a stepping stone to developing more realistic models and analyses, and outline several appealing extensions below.

**Beyond midpoint grading.** In our model, we assume that if all the scores from an interval $[\ell,u]$ are mapped to the same grade, then they are effectively mapped to the grade $(\ell + u)/2$. Given a scheme for converting percentages to letter grades, this is one of the most common ways of converting letter grades back to percentages [mid, a,b]. However, sometimes other values within the range $[\ell,u]$ are used [non].

**Non-uniform letter grading.** Our theoretical results are limited to uniform letter grading schemes, which are used less often in practice than non-uniform letter grading schemes; it would be interesting to extend our theoretical results to study such more general grading schemes. More broadly, in our model, the grading scheme maps the score to a grade from $[0,1]$, which allows a student to compare the grade to their true quality, which is also from $[0,1]$. If the grade is instead in a different numerical range, the model can be easily extended by renormalization (for example, grade point averages lie in $[0,4]$, which is often mapped to $[0,1]$ by dividing by 4). How can our model be extended to incorporate truly non-numeric grades (e.g., A, B, etc.) without reverting to a way to map them to numeric grades (e.g., 4, 3.7, etc.).

**Non-linear (de)motivation.** Our model assumes that the increase or decrease in the true quality is linear in the difference between the received grade and true quality. Evidence from prospect theory suggests that motivational effects from positive outcomes are typically concave (diminishing rewards) while demotivational effects from negative outcomes are typically convex (increasing losses) [Kahneman and Tversky, 1979]. It would be interesting to extend our theoretical results to such nonlinear (de)motivational effects.

References


Appendix

A Intuition Regarding \( D_{\text{same}} \) vs \( D_{\text{opp}} \) & Single-Peakedness

(a) With a small value of \( \gamma \), one can see that within the grade interval \([70, 80]\) containing the true quality \( q = 73 \), the probability of the score being on the same side of the midpoint as the true quality (i.e., in \([70, 75]\)) is significantly higher than the probability of it being on the opposite side of the midpoint (i.e., in \([75, 80]\)). The former region contributes to \( D_{\text{same}} \) while the latter contributes to \( D_{\text{opp}} \). Their difference is the most pronounced when the true quality is near the interval endpoints (e.g., \( q \approx 70, 80 \)) and gradually vanishes when it is near the midpoint (e.g., \( q \approx 75 \)). In expectation over the true quality, one can still expect \( \Pr[(q, s) \in D_{\text{same}}] \) to be sufficiently higher than \( \Pr[(q, s) \in D_{\text{opp}}] \), satisfying the conditions in Theorem 3 and Theorem 4.

(b) Due to single-peakedness of the score distribution, the expected score in any interval lower than the interval containing the true quality \( q = 73 \) is at least its midpoint (e.g., the expected score subject to the score being in \([60, 70]\) is at least 65). In contrast, the expected score in any interval higher than the interval containing the true quality \( q = 73 \) is at most its midpoint (e.g., the expected score subject to the score being in \([80, 90]\) is at most 85). This observation is used at the end of the proof of Theorem 2.

Figure 3: Both figures show the probability density function of the score distribution \( S(q) \) when the true quality is \( q = 73 \). The distribution is a truncated normal distribution with mean \( q = 73 \), and standard deviation \( \gamma = 1.7 \) (top figure) and \( \gamma = 6 \) (bottom figure). The top figure conveys the intuition behind the conditions in Theorem 3 and Theorem 4, which assume \( \Pr[(q, s) \in D_{\text{same}}] \) to be sufficiently higher than \( \Pr[(q, s) \in D_{\text{opp}}] \). The bottom figure conveys the intuition behind the observation used at the end of the proof of Theorem 2.
B Missing Proofs

B.1 Useful Lemmas

Before we dive into the missing proofs, we state two implications of the integral Chebyshev inequality (Lemma 1). The following inequality is obtained by substituting \( f(x) = x \) (and thus, \( \frac{1}{b-a} \int_a^b f(x) \, dx = \frac{a+b}{2} \)) into Lemma 1.

**Lemma 2.** If \( g : [a, b] \to \mathbb{R}_{\geq 0} \) is a non-increasing function, then we have

\[
\int_a^b x g(x) \, dx \leq \frac{a+b}{2} \cdot \int_a^b g(x) \, dx,
\]

and the inequality is reversed if \( g \) is a non-decreasing function.

If \( g \) is a probability density function over \([a, b]\), then

\[
\int_a^b g(x) \, dx = 1,
\]

yielding the following (quite natural) implication.

**Lemma 3.** Let \( X \) be a random variable over \([a, b]\) with a non-increasing probability density function \( g : [a, b] \to \mathbb{R}_{\geq 0} \). Then,

\[
\mathbb{E}[X] \leq \frac{a+b}{2},
\]

and the inequality is reversed if \( g \) is non-decreasing.

Finally, we use the following strengthening of the integral Chebyshev inequality when one of the functions is linear and the other is concave non-increasing.

**Lemma 4.** Let \( g : [a, b] \to \mathbb{R}_{\geq 0} \) be a concave function with \( g(b) = 0 \). Then, we have

\[
\int_a^b (b - x) g(x) \, dx \leq \frac{2(b-a)}{3} \int_a^b g(x) \, dx.
\]

**Proof.** Due to concavity of \( g \), we have

\[
\int_x^b g(t) \, dt \geq \frac{1}{2} (b-x) g(x).
\]

Hence, we have

\[
\int_a^b \frac{1}{2} (b-x) g(x) \, dx \geq \int_x^b \int_{t=x}^b g(t) \, dt \, dx
\]

\[
= \int_{t=a}^b \int_{x=a}^t g(t) \, dx \, dt \quad \text{(Fubini’s theorem)}
\]

\[
= \int_{t=a}^b (t-a) g(t) \, dt
\]

\[
= \int_{x=a}^b (x-a) g(x) \, dx \quad \text{(Change of variable name)}
\]

\[
= \int_a^b (b-a) g(x) \, dx - \int_a^b (b-x) g(x) \, dx.
\]

Rearranging the terms yields the desired inequality. \( \square \)

B.2 Proof of Theorem 3

**Proof.** Given Theorem 2, we only need to show that

\[
\mathbb{E}\left[ |q - s| - |q - ULG_T(s)| \right] \leq 0,
\]

where \( (q, s) \in \mathcal{D} \)

\[
\mathbb{E}\left[ |q - s| - |q - ULG_T(s)| \right] = \Pr[(q, s) \in \mathcal{D}_{\text{same}} \mid (q, s) \in \mathcal{D}] \cdot \mathbb{E}\left[ |q - s| - |q - ULG_T(s)| \right] \mid (q, s) \in \mathcal{D}_{\text{same}}
\]

\[
+ \Pr[(q, s) \in \mathcal{D}_{\text{opp}} \mid (q, s) \in \mathcal{D}] \cdot \mathbb{E}\left[ |q - s| - |q - ULG_T(s)| \right] \mid (q, s) \in \mathcal{D}_{\text{opp}}
\]

\[
\leq 0.
\]

Let us analyze the expected value of \( |q - s| - |q - ULG_T(s)| \) conditioned on both \((q, s) \in \mathcal{D}_{\text{same}}\) and \((q, s) \in \mathcal{D}_{\text{opp}}\) separately.
Analyzing $D_{\text{same}}$. For $k \in \{0, 1, \ldots, T - 1\}$, define $\ell(k) = k \Delta$, $m(k) = (k + 1/2) \Delta$, and $h(k) = (k + 1) \Delta$. These are respectively the lower end, midpoint, and upper end of the $k$-th grade interval under ULG$_T$. Note that

$$D_{\text{same}} = \left\{(q, s) : (\ell(k) \leq q \leq s \leq m(k)) \lor (\ell(k) \leq q \leq s \leq m(k)) \lor (m(k) \leq q \leq s < h(k)) \lor (m(k) \leq q \leq s < h(k)), k \in \{0, 1, \ldots, T - 1\} \right\}.$$ 

Fix an arbitrary $k \in \{0, 1, \ldots, T - 1\}$; write $\ell$, $m$, and $h$ while omitting the fixed $k$ in the argument; and let us analyze the desired expression $[|q - s| - |q - \text{ULG}_T(s)|]$ conditioned on each of the four cases for this fixed $k$ separately. We will derive bounds that will hold regardless of the value of $k$, and, therefore, also conditional on $(q, s) \in D_{\text{same}}$ (i.e., aggregated across all $k$). Note that in each case, we have ULG$_T(q) = \text{ULG}(s) = m$.

1. $\ell \leq q \leq s \leq m$. In this case, $|q - s| - |q - \text{ULG}_T(s)| = s - m$. Note that

$$E[s - m \mid \ell \leq q \leq s \leq m] = \frac{\int_{q=\ell}^{m} \int_{s=q}^{m} f_Q(q) \cdot f_S(s; q) \cdot (s - m) \, ds \, dq}{\Pr[\ell \leq q \leq s \leq m]} = \int_{q=\ell}^{m} 1 \cdot \Pr[s - m \mid q, s \in [q, m]] \cdot \Pr[s \in [q, m] \mid q] \, dq \leq \frac{1}{m - \ell} \cdot \left( \int_{q=\ell}^{m} \Pr[s \in [q, m] \mid q] \, dq \right) \cdot \left( \int_{q=\ell}^{m} \Pr[s \in [q, m] \mid q] \, dq \right) = \frac{1}{\Delta^2} \int_{r=0}^{\Delta/2} \frac{1}{2} \, dr = -\frac{\Delta}{8}.$$ 

Here, the third transition holds because conditioned on a given value of $q$ and on $s \in [q, m]$, the distribution of $s \in [q, m]$ is single-peaked with peak at $q$ (Assumption 2). Hence, $E[s|q, s \in [q, m]] \leq (q + m)/2$. The fourth transition is the integral Chebyshev inequality (Lemma 1), which holds because both $(m - q)/2$ and $\Pr[s \in [q, m] \mid q]$ are non-negative, non-increasing functions of $q$ in $[\ell, m]$ (Assumption 3).

2. $\ell \leq s \leq q \leq m$. In this case, $|q - s| - |q - \text{ULG}_T(s)| = 2q - m - s$. Note that

$$E[2q - m - s \mid \ell \leq s \leq q \leq m] = \frac{\int_{q=\ell}^{m} \int_{s=\ell}^{q} f_Q \cdot f_S(q, s) \cdot (2q - m - s) \, ds \, dq}{\Pr[\ell \leq s \leq q \leq m]} = \int_{s=\ell}^{m} f_S(s) \cdot \Pr[2q - m - s \mid s, q \in [s, m]] \cdot \Pr[q \in [s, m] \mid s] \, ds,$$

Here, we use $f_Q \cdot f_S(q, s) = f_Q(q) \cdot f_S(s; q)$ to denote the joint probability density of $q$ and $s$, and $f_S(s) = \int_{q=0}^{m} f_Q(q) f_S(s; q) \, dq$ to denote the marginal probability density of $s$.

We argue that $E[2q - m - s \mid s, q \in [s, m]] \leq 0$. Intuitively, this is because the posterior distribution of $q \in [s, m]$ conditioned on a fixed value of $s$ and on $q \in [s, m]$, by Assumptions 1 and 2, is single-peaked with peak at $s$. Hence, $E[q \mid s, q \in [s, m]] \leq (s + m)/2$.  

13
Formally, this can be viewed as

\[
\mathbb{E}\left[2q - m - s \mid s, q \in [s, m]\right] = \frac{\int_{q=s}^{m} f(q; s)(2q - m - s) \, dq}{\int_{q=s}^{m} f(q; s) \, dq} \\
\leq \frac{1}{m-s} \cdot \left(\int_{q=s}^{m} f(q; s) \, dq\right) \cdot \left(\int_{q=s}^{m} (2q - m - s) \, dq\right) = 0,
\]

where \( f(q; s) = \frac{f_{\mathcal{Q}}(q) f_{\mathcal{S}}(q; a)}{f_{\mathcal{S}}(q)} \) denotes the probability density of true quality being \( q \) conditioned on the score being \( s \); the second transition is the integral Chebyshev inequality (Lemma 1), which holds because \( f(q; s) \) is a non-increasing function of \( q \) whereas \( 2q - m - s \) is a non-decreasing function of \( q \);⁴ and the final transition holds because the second integral in the numerator is 0.

3. \( m \leq q \leq s < h \). In this case, \(|q - s| - |q - \text{ULG}_T(s)| = m + s - 2q\). Due to the same reasoning as in Case 2, we have that \( \mathbb{E}\left[m + s - 2q \mid m \leq q \leq s < h\right] \leq 0 \).

4. \( m \leq s \leq q < h \). In this case, \(|q - s| - |q - \text{ULG}_T(s)| = m - s\). Due to the same reasoning as in Case 1, we have that \( \mathbb{E}\left[m - s \mid m \leq s \leq q < h\right] \leq -\Delta/8 \).

Let \( p_1, p_2, p_3, p_4 \) respectively denote the total probabilities of the above four cases across all values of \( k \in \{0, 1, \ldots, T - 1\} \), conditioned on \( (q, s) \in \mathcal{D}_{\text{same}} \). Then, \( p_1 + p_2 + p_3 + p_4 = 1 \). Because \( f_{\mathcal{S}}(a; b)/f_{\mathcal{S}}(b; a) \leq \gamma \) for all \( a, b \in [0, 1] \), it follows that \( p_1 \geq p_2/\gamma \) and \( p_4 \geq p_3/\gamma \). Hence, \( p_1 + p_4 \geq (p_2 + p_3)/\gamma \). Using \( p_1 + p_2 + p_3 + p_4 = 1 \), we get \( p_1 + p_4 \geq 1/(\gamma + 1) \).

Combining the analysis from the four cases above, we have

\[
\mathbb{E}\left[|q - s| - |q - \text{ULG}_T(s)| \mid (q, s) \in \mathcal{D}_{\text{same}}\right] \leq -(p_1 + p_4) \cdot \frac{\Delta}{8} \leq -\frac{\Delta}{8(\gamma + 1)}.
\]

(5)

Analyzing \( \mathcal{D}_{\text{opp}} \). Note that

\( \mathcal{D}_{\text{opp}} = \cup_{k \in \{0, 1, \ldots, T - 1\}} \{(q, s) : (\ell(k) \leq q \leq m(k) \leq s \leq h(k)) \lor (\ell(k) \leq s \leq m(k) \leq q \leq h(k))\} \).

Fix an arbitrary \( k \in \{0, 1, \ldots, T - 1\} \); as before, write \( \ell, m, h \) while omitting the fixed \( k \) in the argument. Once again, we analyze the desired expression \(|q - s| - |q - \text{ULG}_T(s)|\) conditioned on each of the two cases in the above expansion of \( \mathcal{D}_{\text{opp}} \) for this fixed \( k \). We will derive bounds that will hold regardless of the value of \( k \), and, therefore, also conditional on \( (q, s) \in \mathcal{D}_{\text{opp}} \) (i.e., aggregated across all \( k \)). Note that we still have \( \text{ULG}_T(q) = \text{ULG}_T(s) = m \).

1. \( \ell \leq q \leq m \leq s \leq h \): In this case, we have \(|q - s| - |q - \text{ULG}_T(s)| = s - m\). Note that

\[
\mathbb{E}\left[s - m \mid \ell \leq q \leq m \leq s \leq h\right] \leq \Delta/4.
\]

(6)

This is because \( s \in [m, m + \Delta/2] \) and, due to single-peakedness of the score model and \( q \leq m \), it is at most \( m + \Delta/4 \) in expectation.

2. \( \ell \leq s \leq m \leq q \leq h \): In this case, we have \(|q - s| - |q - \text{ULG}_T(s)| = m - s\), and the same reasoning as above shows that

\[
\mathbb{E}\left[m - s \mid \ell \leq s \leq m \leq q \leq h\right] \leq \Delta/4.
\]

(7)

⁴To see why \( f(q; s) = \frac{f_{\mathcal{Q}}(q) f_{\mathcal{S}}(q; a)}{f_{\mathcal{S}}(q)} \) is non-increasing in \( q \), note that the denominator does not depend on \( q \) whereas the numerator is equal to \( f_{\mathcal{S}}(s; q) \) (Assumption 1), which is non-increasing in \( q \) (Assumption 3).

⁵Technically, integral Chebyshev inequality requires non-negative functions, and \( 2q - (m + s) \) can be negative when \( q < (m + s)/2 \). However, one can equivalently separate out the \(-(m + s)\) term, apply the integral Chebyshev inequality to \( 2q \), and recombine with the \(-(m + s)\) term to achieve the same conclusion.
Combining Equations (6) and (7) and aggregating over all \(k \in \{0, 1, \ldots, T - 1\}\), we get that
\[
\mathbb{E}\left[|q - s| - |q - \text{ULG}_T(s)| \mid (q, s) \in D^{\text{opp}}\right] \leq \Delta/4. \tag{8}
\]
Finally, combining Equations (5) and (8), we have that
\[
\mathbb{E}\left[|q - s| - |q - \text{ULG}_T(s)| \mid (q, s) \in D\right] \\
\leq \Pr\left[(q, s) \in D^{\text{same}} \mid (q, s) \in D\right] \cdot \left(-\frac{\Delta}{8(\gamma + 1)}\right) + \Pr\left[(q, s) \in D^{\text{opp}} \mid (q, s) \in D\right] \cdot \frac{\Delta}{4} \leq 0,
\]
where the final transition holds because \(\Pr\left[(q, s) \in D^{\text{same}}\right] \geq 2(\gamma + 1) \cdot \Pr\left[(q, s) \in D^{\text{same}}\right]
\)
(Assumption 4), which is equivalent to
\[
\Pr\left[(q, s) \in D^{\text{same}} \mid (q, s) \in D\right] \geq 2(\gamma + 1) \cdot \Pr\left[(q, s) \in D^{\text{same}} \mid (q, s) \in D\right].
\]
This completes the proof. \(\square\)

### B.3 Proof of Theorem 4

**Proof.** As in the proof of Theorem 3, note that given Theorem 2, we only need to prove
\[
\mathbb{E}\left[|q - s| - |q - \text{ULG}_T(s)| \mid (q, s) \in D\right] \\
= \Pr\left[(q, s) \in D^{\text{same}} \mid (q, s) \in D\right] \cdot \mathbb{E}\left[|q - s| - |q - \text{ULG}_T(s)| \mid (q, s) \in D^{\text{same}}\right] \\
+ \Pr\left[(q, s) \in D^{\text{opp}} \mid (q, s) \in D\right] \cdot \mathbb{E}\left[|q - s| - |q - \text{ULG}_T(s)| \mid (q, s) \in D^{\text{opp}}\right] \\
\leq 0. \tag{9}
\]
In the proof of Theorem 3, we analyzed the expected value of \(|q - s| - |q - \text{ULG}_T(s)|\) conditioned on both \((q, s) \in D^{\text{same}}\) and \((q, s) \in D^{\text{opp}}\) separately: the former was shown to be at most \(-\frac{\Delta}{8(\gamma + 1)}\)
whereas the latter was shown to be at most \(\frac{\Delta}{4}\), yielding the desired Equation (9) when \(\Pr\left[(q, s) \in D^{\text{same}}\right] \geq 2(\gamma + 1) \cdot \Pr\left[(q, s) \in D^{\text{same}}\right]
\).

With the strong symmetry assumption, we improve the former upper bound to \(-\frac{\Delta}{12}\), which improves the sufficient condition to \(\Pr\left[(q, s) \in D^{\text{same}}\right] \geq 3 \cdot \Pr\left[(q, s) \in D^{\text{opp}}\right]\). That is, our goal is to prove
\[
\mathbb{E}\left[|q - s| - |q - \text{ULG}_T(s)| \mid (q, s) \in D^{\text{same}}\right] \leq -\frac{\Delta}{12}.
\]
Note that \(D^{\text{same}} = \bigcup_{k \in \{0, 1, \ldots, T - 1\}} D^{\text{same}}_k\), where \(D^{\text{same}}_k = D^{\text{same}} \cap [k\Delta, (k + 1)\Delta]^2\). We show that \(\mathbb{E}\left[|q - s| - |q - \text{ULG}_T(s)| \mid (q, s) \in D^{\text{same}}_k\right] \leq -\frac{\Delta}{12}\) for all \(k \in \{0, 1, \ldots, T - 1\}\), which implies the desired result. Fix any \(k \in \{0, 1, \ldots, T - 1\}\), and write \(\ell = k\Delta, m = (k + 1/2)\Delta,\) and \(h = (k + 1)\Delta\).

Let us further partition \(D^{\text{same}}_k\) as \(D^{\text{same}}_{k,\text{low}} \cup D^{\text{same}}_{k,\text{high}}\), where \(D^{\text{same}}_{k,\text{low}} = \{(q, s) : \ell \leq q, s < m\}\) (both the true quality and the score are lower than the midpoint) and \(D^{\text{same}}_{k,\text{high}} = \{(q, s) : m \leq q, s < h\}\) (both the true quality and the score are at least as high as the midpoint). Crucially, we note that
\[
\mathbb{E}\left[|q - s| - |q - m| \mid (q, s) \in D^{\text{same}}_{k,\text{low}}\right] = \mathbb{E}\left[|q - s| - |q - m| \mid (q, s) \in D^{\text{same}}_{k,\text{high}}\right].
\]
This is because the transformation \((q, s) \mapsto (q', s')\), where \(q' = m + (m - q)\) and \(s' = m + (m - s)\), is a bijection mapping each point \((q, s) \in D^{\text{same}}_k\) to a point \((q', s') \in D^{\text{same}}_{k,\text{high}}\) with \(|q - s| - |q - m| = |q' - s'| - |q' - m|\); the last observation relies on \(Q\) being a uniform distribution (Assumption 1) and \(\mathcal{S}\) being strongly symmetric (Assumption 2).

Hence, it is sufficient to show that
\[
\mathbb{E}\left[|q - s| - |q - m| \mid (q, s) \in D^{\text{same}}_{k,\text{low}}\right] \leq -\frac{\Delta}{12}.
\]
Next, we further partition \( D_{\text{same}} \) as \( D_{\text{same}} = D_{\text{same}} \cup D_{\text{same, dec}} \), where \( D_{\text{same, low, inc}} = \{(q, s) : \ell \leq q \leq s < m \} \) (the score is at least as much as the true quality) and \( D_{\text{same, low, dec}} = \{(q, s) : \ell \leq s \leq q < m \} \) (the score is at most as much as the true quality). Note that

\[
\begin{align*}
\mathbb{E}[q - s| |q - m| |(q, s) \in D_{\text{same, low, inc}}] &= \Pr[(q, s) \in D_{\text{same, low, inc}} | (q, s) \in D_{\text{same, low, dec}}] \cdot \mathbb{E}[(q - s| |q - m| |(q, s) \in D_{\text{same, low, inc}}] \\
+ \Pr[(q, s) \in D_{\text{same, low, dec}} | (q, s) \in D_{\text{same, low, dec}}] \cdot \mathbb{E}[(q - s| |q - m| |(q, s) \in D_{\text{same, low, dec}}] .
\end{align*}
\]

First, we argue that

\[
\Pr[(q, s) \in D_{\text{same, low, inc}} | (q, s) \in D_{\text{same, low, dec}}] = \Pr[(q, s) \in D_{\text{same, low, dec}} | (q, s) \in D_{\text{same, low, dec}}] = \frac{1}{2}.
\]

This follows by noting the bijection from \( D_{\text{same, low, inc}} \) to \( D_{\text{same, low, dec}} \) given by \((q, s) \rightarrow (q', s')\), where \( q' = m - (q - \ell) \) and \( s' = m - (s - \ell) \); due to strong symmetry of \( S \) and \(|q - s| = |q' - s'|\), we have \( f_{Q \times S}(q, s) = f_{Q \times S}(q', s')\).

Next, recall that in the proof of Theorem 3 (Case 2 in the analysis of \( D_{\text{same}} \)), we had already argued

\[
\mathbb{E}[q - s| |q - m| |(q, s) \in D_{\text{same, low, dec}}] = \mathbb{E}[2q - m| \ell \leq s \leq q < m] \leq 0.
\]

Hence, we have

\[
\mathbb{E}[q - s| |q - m| |(q, s) \in D_{\text{same, low, inc}}] \leq \frac{1}{2} \cdot \mathbb{E}[q - s| |q - m| |(q, s) \in D_{\text{same, low, dec}}],
\]

which means it is sufficient to argue

\[
\mathbb{E}[q - s| |q - m| |(q, s) \in D_{\text{same, low, inc}}] = \mathbb{E}[s - m| \ell \leq q \leq s < m] \leq -\frac{\Delta}{6}.
\]

Note that

\[
\begin{align*}
\mathbb{E}[s - m| \ell \leq q \leq s < m] &= -\frac{\int_{q=\ell}^{m} \int_{s=q}^{m}(m-s)f_{S}(s; q) \, ds \, dq}{\Pr[\ell \leq q \leq s < m]} \quad (Q \text{ is uniform}) \\
&\leq -\frac{\int_{q=\ell}^{m} \int_{s=q}^{m}(m-s) \, ds \cdot \left(\int_{s=q}^{m} f_{S}(s; q) \, ds\right) \, dq}{\Pr[\ell \leq q \leq s < m]} \quad (\text{Lemma 1}) \\
&= -\frac{1}{2} \frac{\int_{q=\ell}^{m}(m-q) \Pr[s \in [q, m]] \, dq}{\Pr[\ell \leq q \leq s < m]} \\
&\leq -\frac{1}{2} \frac{\int_{q=\ell}^{m} \Pr[s \in [q, m]] \, dq}{\Pr[\ell \leq q \leq s < m]} \quad (\text{Lemma 4}) \\
&= -\frac{1}{2} \frac{\int_{q=\ell}^{m} \Pr[\ell \leq q \leq s < m]}{\Pr[\ell \leq q \leq s < m]} = -\frac{\Delta}{6},
\end{align*}
\]

as needed. Here, in the application of Lemma 4 in the fourth transition, we use the fact that \( g(q) = \Pr[s \in [q, m]] = \int_{q=\ell}^{m} f_{S}(s; q) \, ds \) is a concave function and \( g(m) = 0 \). To see concavity, note that strong symmetry of \( S \) means that there is a distribution with probability density \( z \) such that \( f_{S}(s; q) = z(s - q) \). Then, \( g(q) = \int_{s=q}^{m} z(s - q) \, ds = \int_{x=q}^{m-q} z(x) \, dx \). Hence, \( g'(q) = -z(m - q) \) and \( g''(q) = z'(m - q) \). Due to the single-peakedness of \( S \), we have that \( z'(x) \leq 0 \) for all \( x \geq 0 \), so \( g''(q) \leq 0 \), which proves concavity of \( g \). \( \square \)
C Additional Experimental Results

In our experiments, we compared numerical scoring to uniform letter grading schemes with $T \in \{4, 8, 12, 16, 20\}$ grades. In the results presented in Section 4, we presented results that show the impact of two parameters, the number of evaluations $r$ and the motivation coefficient $\alpha_m$, when one of them is varied while keeping the other fixed.

Here, we present additional experimental results, which show the impact of varying other parameters such as the mean true quality $\mu$ (Figure 4), the standard deviation of the true quality prior $\sigma$ (Figure 5), and the standard deviation of the score distribution $\gamma$ (Figure 6). When varying each parameter, we fix the default values of all other parameters mentioned in Section 4. For each variation, we measure the impact on all three metrics described in Section 4: the final score discrepancy, the average score discrepancy, and the inversion probability.

Overall, the mean true quality $\mu$ has little impact on the performance of and the comparison between the different grading schemes. Similarly, the standard deviation $\sigma$ of the true quality prior also does not significantly affect the final and average score discrepancies of most of the grading schemes, but somewhat strikingly, it has a dramatic impact on the final and average score discrepancies of $\text{ULG}_4$ (uniform letter grading with 4 grades). The inversion probability generally decreases as $\sigma$ increases: inversions naturally become rarer when the students have very different true qualities. The impact of varying the standard deviation $\gamma$ of the score distribution is similar to that of varying $\sigma$, except higher $\gamma$ leads to higher inversion probability: when the scores of students can easily be far from their true qualities, inversions naturally become more likely.

There are a couple of striking observations that hold true across all of these experiments.

First, the comparison between numerical scoring and uniform letter grading in terms of the final and average score discrepancies is dependent on the relation between the motivation and demotivation coefficients: numerical scoring is better when $\alpha_m < \alpha_d$ whereas uniform letter grading is better when $\alpha_m > \alpha_d$. This is consistent with our theoretical results (although, note that this can flip when the number of evaluations is increased; see Figure 1).

Second, the comparison between the grading schemes in terms of the inversion probability is not dependent on the relation between $\alpha_m$ and $\alpha_d$: numerical scoring almost always leads to a lower inversion probability than uniform letter grading (and under uniform letter grading, more grades is better than fewer grades). This generally makes sense because letter grading schemes are coarse, giving the same grade to students with different scores (and, thus, likely with different true qualities). That said, this is still not a trivial observation because, as demonstrated via examples in the introduction, letter grading can sometimes also reduce the inversion probability by eliminating the noisy variation of scores between students of equal true qualities. We view this our experimental insight as advocating the use of numerical scoring in the interest of fairness to the students (as measured by reduced inversion probability).
Figure 4: Final score discrepancy (row 1), average score discrepancy (row 2) and inversion probability (row 3) when varying $\mu$, with $r = 2$, $\sigma = 12$, $\gamma = 1.5$, and $\alpha_d = 0.5$. 

(a) $\alpha_m = 0.2$

(b) $\alpha_m = 0.8$

(c) $\alpha_m = 0.2$

(d) $\alpha_m = 0.8$

(e) $\alpha_m = 0.2$

(f) $\alpha_m = 0.8$
Figure 5: Final score discrepancy (row 1), average score discrepancy (row 2) and inversion probability (row 3) when varying $\sigma$, with $r = 2$, $\mu = 65$, $\gamma = 1.5$, and $\alpha_d = 0.5$. 
Figure 6: Final score discrepancy (row 1), average score discrepancy (row 2) and inversion probability (row 3) when varying $\gamma$, with $r = 2$, $\mu = 65$, $\sigma = 12$, and $\alpha_d = 0.5$. 