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

Evi Micha[1] , **Shreyas Sekar**[2,3] , **Nisarg Shah**[1]
[1]Department of Computer Science, University of Toronto
[2]Rotman School of Management, University of Toronto
[3]Department of Management, University of Toronto Scarborough
{emicha, nisarg}@cs.toronto.edu, shreyas.sekar@rotman.utoronto.ca

Abstract

We study letter grading schemes, which are routinely employed for evaluating student performance. Typically, a numerical score obtained via 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 studying the (de)motivational effects of such grading on the students and, consequently, on their performance in future evaluations. We use the model to compare uniform letter grading schemes, in which the range of scores is divided into equal-length parts that are mapped to the letter grades, 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 find cases where the converse occurs.

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 grades.[1] As far as we are aware, this work is among the first to look at the problem of designing a grading scheme with the explicit objective of 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 (i.e., in expectation perform as well as) a specific score or performance level based on their intrinsic ability. We refer to this anchor as the intrinsic quality.
2. **(De)Motivation**: When the students' actual score falls above (below) their intrinsic quality, they get (de)motivated and subsequently, their expectation increases (decreases) for future tests. This is a phenomenon that has been widely noticed in practice [Deci *et al.*, 1999; Dev, 1997; Cameron and Pierce, 1994].

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 a 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 a student with an intrinsic quality of $q_1 = 85$. Suppose that the student scores $s_1 = 81$ in the midterm exam. Disclosing this numeric score may demotivate the student, which may reduce her effective intrinsic quality for the final exam. This adverse effect may be prevented if a (coarser) letter grading scheme is used, in which (say) all students (including the student under consideration) whose scores lie in $[80, 90]$ are assigned a letter grade of $A-$. However, consider another student whose intrinsic quality is $q_2 = 91$ and whose midterm score is $s_2 = 89$. Receiving the same letter grade $A-$ as everyone who scored in $[80, 90]$ may be more demotivating to her than receiving her numerical score of $s_2 = 89$. Hence, the overall effect of using a letter grading scheme remains unclear.

[1]We 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.

There is another subtle issue to be considered. While the comparison made by a student between her numerical score and intrinsic true quality is straightforward,[2] it is not obvious how a student should compare her intrinsic true quality to a letter grade received (such as $A-$). This depends on how the student perceives the letter grade. To that end, we use a scheme for mapping letter grades back to representative numerical scores: each letter grade is mapped to the *midpoint* of the interval containing all the scores that were mapped to that letter grade. For example, if all scores in the range $[80, 90]$ are mapped to the letter grade $A-$, then $A-$ is mapped back to (i.e., considered worth) a score of 85, which is what any student receiving $A-$ would compare to her intrinsic quality.

This *midpoint scheme* has three in-built advantages. First, it reflects how the letter grades may truly be perceived in the outside world (and thus by the students) as it is actually used in the real world [University of Western Ontario, 2022; Victoria University of Wellington, 2022]. Second, the association of a letter grade to the midpoint of its score range accurately conveys the (average) performance of a student receiving that letter grade. Third, in the absence of any structure on how the grades are perceived, the question we ask in this work — which letter grading schemes would lead to the maximum average student performance? — would have a trivial and rather unsatisfactory answer: assign all students a grade worth 100 to maximally motivate them. The midpoint scheme makes this impossible: if the same grade (say $A+$) was assigned to all the students with scores anywhere in $[0, 100]$, it would only be worth 50.

Building on the ideas presented in this example, we develop a framework to compare various grading systems in an environment with *sequential testing*. This includes evaluations within a course, e.g., a midterm followed by a final exam, but also grading across related courses, e.g., a student taking Calculus 101 followed by Calculus 102. Since a student's intrinsic quality increases after a test if the grade received is higher than her intrinsic quality and decreases otherwise, our aim is

> *to compare different grading schemes and choose the one that provides a higher quality improvement (or a lower quality degradation).*

Our results. In this work, we compare the numerical scoring scheme, where the student learns her exact score in an evaluation, to uniform letter grading schemes, where the interval of scores is partitioned into T equal-length intervals mapping to different letter grades (and each interval is represented by its midpoint). While uniform letter grading is not completely realistic, we view our work as a starting point for the curiously unaddressed problem of quantitatively optimizing letter grading schemes and a stepping stone for future work to build on. That said, we note that real-world letter grading schemes (at least those used in North American universities) are close to uniform, once a very large interval mapped to the failing grade and a somewhat large interval mapped to very top grade are omitted. Since very few students fall in these two intervals, this omission does not significantly affect the overall analysis.

First, we theoretically study the case where two sequential evaluations take place, such as midterm and final exam. We show that under natural conditions, numerical scoring and all uniform letter grading schemes have equal performance when the motivational and demotivational effects are equally strong, and otherwise, either numerical scoring outperforms all uniform letter grading schemes or the opposite happens. Analytically identifying when each scheme outperforms the other turns out to be far from obvious and subtly dependent on properties of the distributions of intrinsic true qualities and scores, even for this limited setting. Using carefully constructed bijections between students, we are able to identify additional conditions under which numerical scoring outperforms all uniform letter grading schemes when the demotivational effect is stronger than the motivational effect, and the opposite happens when the demotivational effect is weaker than the motivational effect. Since there is significant evidence that negative events have a greater impact than positive events [Baumeister *et al.*, 2001; Coleman *et al.*, 1987], we expect the demotivational effect to be stronger than the motivational effect; thus, our results are in favour of numerical scoring.

Next, we empirically compare numerical scoring to uniform letter grading schemes. Under two sequential evaluations, we observe that numerical scoring continues to outperform uniform letter grading when the demotivational effect is stronger (and the opposite continues to hold when the motivational effect is stronger), even under more realistic conditions than in our theoretical analysis, such as when the true qualities of the students follow a (truncated) normal distribution. However, surprisingly, when more than two evaluations take place, the effect is reversed. Even after just six sequential evaluations, uniform letter grading begins to outperform numerical scoring when the demotivational effect is stronger (and the opposite holds when the motivational effect is stronger). In the intermediate stage between these two regimes, there is another surprising effect: with four sequential evaluations, numerical scoring outperforms uniform letter grading regardless of which effect is stronger!

Our results indicate that the choice of the grading scheme depends on the application at hand: with fewer evaluations (e.g., courses with just a few tests or shorter education programs with just a few semesters), numerical scoring may be better, while with many evaluations (e.g., courses with weekly tests or longer education programs), uniform letter grading may be better. At a high level, although our work draws on literature from fields such as economics and psychology, it provides a fundamental perspective on the question of student grading within the framework of multi-agent systems, i.e., where each student is modeled as an agent whose behavior depends on the decisions made by the system. Our results open up the possibility of designing grading systems that are easy to implement, approximately-optimal, and take into account students' incentives.

Related work. There is a rich literature on comparing grading schemes using various objectives. However, to the best of our knowledge, none of these papers study the objective of

[2] This assumes that students is aware of their own intrinsic true qualities, but it may suffice for them to have noisy estimates.

improving student quality that we focus on.

Several works have studied, both theoretically and empirically, how the effort exerted by students for an evaluation depends on the grading scheme to be used [Paredes, 2017; Brownback, 2018; Main and Ost, 2014; Czibor *et al.*, 2020]. For example, when using pass/fail grading, a student may try hard enough to pass (with high probability), but not any harder. Our work is orthogonal to this: we focus on effect of the outcome of one evaluation on the student motivation in *subsequent* evaluations.

Another related work is that of Sikora [2015], who also compares grading schemes, but his goal is to study the trade-off between conveying the most information about the student's true quality and minimizing noise due to factors unrelated to the true quality, not the (de)motivational effects of the grading scheme in subsequent evaluations. In our work, the task of keeping the grades "consistent" with the actual performance is indirectly performed by the midpoint scheme.

Rohe *et al.* [2006] and Bloodgood *et al.* [2020] also study how the grading scheme used may impact students' psychological well-being and stress levels, but do not focus on the impact of this in subsequent evaluations.

2 Model

Define $[k] = \{1, \dots, k\}$ 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 $B : [0, 1] \to [0, 1]$ that maps the score to a grade.

Letter grading. A *letter grading* scheme $B_{\vec{c}}$ is specified by a vector $\vec{c} = (c_0 = 0, c_1, \dots, c_{T-1}, c_T = 1)$, for some $T \in \mathbb{N}$ (referred to as the number of grades) and $c_i \geqslant c_{i-1}$ for all $i \in [T]$, and is given by $B_{\vec{c}}(s) = \frac{c_{i-1}+c_i}{2}$ 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 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 $s \in [0, 1)$,[3]

$$\mathrm{ULG}_T(s) = (\lfloor s/\Delta \rfloor + 1/2) \cdot \Delta.$$

For instance, 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. We restrict our focus to uniform grading schemes for two reasons: a) it is straightforward and easy to implement in practice; b) given that different institutions following different grading schemes, this allows us to broadly compare letter and number grading without getting lost in the minutiae. Further our assumption that each letter grade maps to the midpoint of an interval is common practice across universities [University of Western Ontario, 2022; Victoria University of Wellington, 2022] as well as the literature [McEwan *et al.*, 2021; Nisbet, 1975]. More generally, it is consistent with the practice of assigning a score or grade-point to each letter grade.

Numerical scoring. We will compare (uniform) letter grading to *numerical scoring* (NS), given by $\mathrm{NS}(s) = s$ 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 $T \to \infty$ 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

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

We refer to $\alpha_m, \alpha_d \in \mathbb{R}_{\geqslant 0}$ 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 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.[4] Our choice of a linear model for demotivation follows from studies showing that student performance is linearly dependent on both external [Christensen and Menzel, 1998] and internal stimuli [Latham and Locke, 2007]. Additionally, even when the actual behaviour is more complex, our model serves as a first-order approximation when $(B(s) - q)$ is small.

Goal. Intuitively, we are interested in choosing grading schemes that achieve a higher increase (or a lower decrease) in the average student quality. Thus, we define the *performance* of a grading scheme B as:

$$\mathrm{perf}(B) \triangleq \mathbb{E}_{q\sim\mathcal{Q}, s\sim\mathcal{S}(q)}[q' - q]$$

where $q' = q + h(q, B(s))$. Due to linearity of expectation,

$$\mathrm{perf}(B) = \mathbb{E}_{q\sim\mathcal{Q}, s\sim\mathcal{S}(q)}[q' - q] = \mathbb{E}_{q\sim\mathcal{Q}, s\sim\mathcal{S}(q)}[h(q, B(s))].$$

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

[4] In principle, one can also use larger coefficients and truncate q' to lie in $[0, 1]$.

Thus, we compare $\mathbb{E}_{q\sim\mathcal{Q},s\sim\mathcal{S}(q)}[h(q,B(s))]$ under numerical scoring and uniform letter grading. Hereinafter, we omit $q \sim \mathcal{Q}$ and $s \sim \mathcal{S}(q)$ from an expression of expectation, whenever it is clear from the context.

Note that for our theoretical analysis, we focus on the case of two evaluations. Later, we empirically study the case of more than two evaluations.

3 Theoretical Results

In this section, we derive theoretical results for the performance of uniform letter grading schemes and numerical scoring, when students participate in two sequential evaluations and identify conditions under which numerical scoring outperforms every 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 $\mathcal{Q}$ and the score model $\mathcal{S}$ are *jointly symmetric* if $f_{\mathcal{Q}}(q) \cdot f_{\mathcal{S}}(s;q) = f_{\mathcal{Q}}(1-q) \cdot f_{\mathcal{S}}(1-s;1-q)$ 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 $\mathcal{S}(q)$ should be the mirror image of the score distribution $\mathcal{S}(1-q)$. 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 need $f_{\mathcal{S}}(s=0.4;q=0.5) = f_{\mathcal{S}}(s=0.6;q=0.5)$.

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 $\mathcal{Q}$ and the score model $\mathcal{S}$ are jointly symmetric, and the grading scheme B is symmetric, then we have*

$$\mathrm{perf}(B) = \frac{\alpha_m - \alpha_d}{2} \cdot \mathbb{E}_{q\sim\mathcal{Q},s\sim\mathcal{S}(q)}\Big[|q - B(s)|\Big]. \quad (1)$$

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

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

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. Hence, we have that $\mathbb{E}\Big[h(q,B(s)) + h(1-q,B(1-s))\Big] = (\alpha_m - \alpha_d) \cdot \mathbb{E}\Big[|q - B(s)|\Big]$. Plugging this into Equation (2), we get the result. □

Corollary 1. *Assume that the true quality prior $\mathcal{Q}$ and the score model $\mathcal{S}$ are jointly symmetric. Then, all symmetric grading schemes have equal performance if $\alpha_m = \alpha_d$. Further, if $\alpha_m \neq \alpha_d$, 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 if $\alpha_m > \alpha_d$, and the converse holds if $\alpha_m < \alpha_d$.*
2. *Uniform letter grading with T grades is at least as good as numerical scoring if $\alpha_m < \alpha_d$, and the converse holds if $\alpha_m > \alpha_d$.*

Proof. The first claim regarding $\alpha_m = \alpha_d$ follows immediately from Equation (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 $\mathbb{E}[|q - \mathrm{NS}(s)| - |q - \mathrm{ULG}_T(s)|]$, 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 (Ex-Ante Single-Peaked Score Model)**.** We say that the score model $\mathcal{S}$ is *ex-ante single-peaked* if, for every $q \in [0,1]$, $f_{\mathcal{S}}(\cdot;q)$ is single-peaked with the peak at q, i.e., $f_{\mathcal{S}}(s;q) \leqslant f_{\mathcal{S}}(s';q)$ for all $s \leqslant s' \leqslant q$ and $s \geqslant s' \geqslant q$.

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

For a fixed T, we also denote with $\mathcal{D}$ the set of all pairs of true qualities and scores that belong to the same letter grade interval, i.e., $\mathcal{D} = \{(q,s) : \mathrm{ULG}_T(q) = \mathrm{ULG}_T(s)\}$. For example, if $T = 10$, $(q = 0.51, s = 0.59) \in \mathcal{D}$ but $(q = 0.51, s' = 0.49) \notin \mathcal{D}$.

Theorem 2. *Fix any $T \in \mathbb{N}$. Assume that the true quality prior $\mathcal{Q}$ and the score model $\mathcal{S}$ satisfy the following.*

1. *$\mathcal{Q}$ and $\mathcal{S}$ are jointly symmetric;*
2. *$\mathcal{S}$ is ex-ante single-peaked; and*
3. $\mathbb{E}\Big[|q-s| \;\Big|\; (q,s) \in \mathcal{D}\Big] \leqslant \mathbb{E}\Big[|q - \mathrm{ULG}_T(s)| \;\Big|\; (q,s) \in \mathcal{D}\Big]$.

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

Before diving into the proof, let us make a remark regarding the third technical condition in Theorem 2. 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 sketch. Given Theorem 1, we simply need to show that $\mathbb{E}\Big[|q-s|\Big] \leqslant \mathbb{E}\Big[|q-\mathrm{ULG}_T(s)|\Big]$. We already assume that this holds conditioned on $(q,s) \in \mathcal{D}$. Hence, we only need to show that it also holds conditioned on $(q,s) \notin \mathcal{D}$. We show this given the additional single-peakedness property. In fact, we show that conditioned on $(q,s) \notin \mathcal{D}$, the desired equation actually holds for every $q \in [0,1]$, and, thus, in expectation over $q \sim \mathcal{Q}$ too. To see why this is true, fix any $q \in [0,1]$ and consider two cases based on whether the letter grade mapping for score s (i.e., $\mathrm{ULG}_T(s)$) is smaller or larger than the corresponding grade for q (i.e., $\mathrm{ULG}_T(q)$).

In the first case where $\mathrm{ULG}_T(s) < \mathrm{ULG}_T(q)$, the single-peakedness property implies that $|q-s| = q-s \leqslant q - \mathrm{ULG}_T(s)$ in expectation over s. Conversely, when $\mathrm{ULG}_T(s) > \mathrm{ULG}_T(q)$, one can infer that $|q-s| = s-q \leqslant \mathrm{ULG}_T(s) - q = |q - \mathrm{ULG}_T(s)|$ in expectation over s, based on the same argument. Putting these two cases together, we get the third statement in the theorem. A formal proof can be found in the full version. □

In Theorem 2, we argued that single-peakedness of $\mathcal{S}$ establishes the desired inequality of $\mathbb{E}\Big[|q-s|\Big] \leqslant \mathbb{E}\Big[|q-\mathrm{ULG}_T(s)|\Big]$ at least conditioned on $(q,s) \notin \mathcal{D}$, leaving only the case of $(q,s) \in \mathcal{D}$, which was stated as an assumption in. Next, we show that if the true quality prior $\mathcal{Q}$ is uniform over $[0,1]$, and it satisfies two natural assumptions then the desired inequality also holds conditioned on $(q,s) \in \mathcal{D}$.

Definition 4 (Ex-Post Single-Peaked Score Model)**.** We say that the score model $\mathcal{S}$ is *ex-post single-peaked* if, for every $s \in [0,1]$, $f_{\mathcal{S}}(s;\cdot)$ is single-peaked with the peak at s, i.e., $f_{\mathcal{S}}(s;q) \leqslant f_{\mathcal{S}}(s;q')$ for all $s \leqslant q' \leqslant q$ and $q \leqslant q' \leqslant s$.

Definition 5 (Probabilistic Single-Dipped Score Model)**.** We say that the score model $\mathcal{S}$ is *probabilistic single-dipped* if, for every $x \in [0,1]$, $\Pr\Big[s \in [q,x] \cup [x,q] \;\Big|\; q\Big]$ (let us call this $p(x,q)$) is single-dipped in q with the dip at $q=x$, i.e., $p(x,q) \leqslant p(x,q')$ for all $x \leqslant q' \leqslant q$ and $q \leqslant q' \leqslant x$.

Before we state the next theorem, we further partition $\mathcal{D}$ into two sub-spaces, $\mathcal{D}^{\mathrm{same}}$ and $\mathcal{D}^{\mathrm{opp}}$, such that $\mathcal{D}^{\mathrm{same}}$ contains the set of all pairs of true qualities and scores such that either both are at most or both are at least the midpoint of their common letter grade interval, i.e.

$$\mathcal{D}^{\mathrm{same}} = \{(q,s) : q,s \leqslant \mathrm{ULG}_T(q) = \mathrm{ULG}_T(s) \\ \vee\, q,s \geqslant \mathrm{ULG}_T(q) = \mathrm{ULG}_T(s)\}$$

and $\mathcal{D}^{\mathrm{opp}} = \mathcal{D} \setminus \mathcal{D}^{\mathrm{same}}$. For example, when $T=10$, $(q=0.54, s=0.51) \in \mathcal{D}^{\mathrm{same}}$, but $(q=0.54, s'=0.56) \in \mathcal{D}^{\mathrm{opp}}$. We are now ready to state the result.

Theorem 3. *Fix arbitrary $T \in \mathbb{N}$. Assume the following regarding the true quality prior $\mathcal{Q}$ and the score model $\mathcal{S}$.*

1. *$\mathcal{Q}$ is uniform over $[0,1]$;*
2. *$\mathcal{Q}$ and $\mathcal{S}$ are jointly symmetric;*
3. *$\mathcal{S}$ is ex-ante and ex-post single-peaked, and probabilistic single-dipped; and*
4. $\Pr\Big[(q,s) \in \mathcal{D}^{\mathrm{same}}\Big] \geqslant 2(\gamma+1) \cdot \Pr\Big[(q,s) \in \mathcal{D}^{\mathrm{opp}}\Big]$, *where* $\gamma = \max_{a,b \in [0,1]} \frac{f_{\mathcal{S}}(a;b)}{f_{\mathcal{S}}(b;a)}$.

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

The proofs of Theorems 3 and 4 are our most intricate proofs. However, due to space constraints, we have deferred them to the full version.[5] Let us now understand the assumptions in Theorem 3. A natural choice of $\mathcal{S}$ under which Assumptions 3 and 4 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 $\Pr[(q,s) \in \mathcal{D}^{\mathrm{same}}] \geqslant 4 \cdot \Pr[(q,s) \in \mathcal{D}^{\mathrm{opp}}]$. More general, from the definitions of $\mathcal{D}^{\mathrm{same}}$ and $\mathcal{D}^{\mathrm{opp}}$, when the variance of the score distribution is sufficiently small, we can expect $\Pr[(q,s) \in \mathcal{D}^{\mathrm{same}}]$ to be much higher than $\Pr[(q,s) \in \mathcal{D}^{\mathrm{opp}}]$. The full version includes a figure providing further intuition.

Ex-ante single-peakedness, ex-post single-peakedness, and probabilistic single-dippedness can be subsumed into a single property that captures a stronger form of symmetry, in which the noise in the score is symmetric and zero-mean.

Definition 6 (Strongly Symmetric Score Model)**.** We say that the score model $\mathcal{S}$ is strongly symmetric if $f_{\mathcal{S}}(s;q) = \ell(|s-q|)$ for some non-increasing function $\ell : \mathbb{R}_{\geqslant 0} \to \mathbb{R}_{\geqslant 0}$.

Under a strongly symmetric score model, we have $\gamma = 1$ in Assumption 4 of Theorem 3, which means a constant of $2(\gamma+1) = 4$ would be needed. However, using different techniques, we can show that even a constant of 3 suffices to obtain the same result under strong symmetry. This broadens the scope to include less concentrated score models.

Theorem 4. *Fix arbitrary $T \in \mathbb{N}$. Let $\mathcal{D}$, $\mathcal{D}^{\mathrm{same}}$, and $\mathcal{D}^{\mathrm{opp}}$ be defined as in Theorem 3. Assume the following regarding the true quality prior $\mathcal{Q}$ and the score model $\mathcal{S}$.*

1. *$\mathcal{Q}$ is uniform over $[0,1]$;*
2. *$\mathcal{S}$ is strongly symmetric; and*
3. $\Pr\Big[(q,s) \in \mathcal{D}^{\mathrm{same}}\Big] \geqslant 3 \cdot \Pr\Big[(q,s) \in \mathcal{D}^{\mathrm{opp}}\Big]$.

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

We remark that in the proof of Theorem 4, we only really need strong symmetry for pairs of true qualities and scores that belong to the same letter grade interval.

[5] Full version: www.cs.toronto.edu/~nisarg/papers/grading.pdf

4 Experiments

In the previous section, we proved that when $\mathcal{Q}$ is uniformly distributed and the variance of the score model is small, we can conclude that the first implication of Corollary 1 holds. In this section, we empirically compare numerical scoring and uniform letter grading while relaxing these assumptions.

First, it is widely believed that students' true qualities, at least in large classes, are normally distributed based on the evidence that "exam scores tend to be normally distributed for well-constructed, norm-referenced, multiple choice tests" [Wedell *et al.*, 1989]. Hence, we empirically study the case where $\mathcal{Q}$ is normally distributed, truncated to $[0, 1]$. We also consider cases where the score is not necessarily concentrated around the true quality. Finally, our analysis was limited to two evaluations; in our experiments, we also consider more than two evaluations. When a student participates in r sequential evaluations, after each evaluation the student compares her "current" true quality to the obtained grade, and experiences (de)motivation that affects her effective true quality in the next evaluation. Formally, for $j \in [r]$, let q_j and s_j denote her effective true quality and score in evaluation j, respectively. Then, $s_j \sim \mathcal{S}(q_j)$ for each $j \in [r]$, and for $j \in [r-1]$, we have:

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

We measure the performance of a grading scheme by comparing the final true quality, q_r, to the initial true quality q_1, which extends the performance measure introduced in preliminaries for two evaluations:

$$\mathrm{perf}_F(B) \triangleq \mathbb{E}_{q\sim\mathcal{Q}, s\sim\mathcal{S}(q)}[q_r - q_1].$$

Data generation. For all the simulations, we compare numerical scoring (NS) to uniform letter grading (ULG_T) with $T \in \{4, 8, 12, 16, 20\}$ grades. We scale the interval of grades to $[0, 100]$ to resemble percentage grades. We simulate $n = 5000$ students (average results are plotted with 95% confidence intervals), where the initial true quality q_1 of each student is drawn i.i.d. from a truncated normal distribution capped to $[0, 100]$, with the underlying normal distribution characterized by mean μ and standard deviation σ. Given a true quality q in an evaluation, the score s is drawn from another truncated normal distribution capped to $[0, 100]$, with the underlying normal distribution characterized by mean q and standard deviation γ.

Results. Figure 1 shows how the final quality improves (or degrades) with respect to the motivation coefficient (top) and the number of evaluations (bottom). In Figure 1a, the motivation coefficient takes values in $\{0, 0.1 \ldots, 0.9, 1\}$, the demotivation coefficient is set to 0.5 and the number of evaluations is set to $r = 2$. We see that when $\alpha_m < \alpha_d$, numerical scoring is better than any uniform letter grading (and uniform letter grading with more grades is better than uniform letter grading with fewer grades), whereas when $\alpha_m > \alpha_d$, the opposite is true. Hence, it seems that the first implication of Corollary 1 continues to hold, even when the true quality is drawn from more realistic distributions. The comparison between uniform letter grading schemes with different numbers of grades is intuitive: uniform letter grading essentially converges to numerical scoring when T goes to infinity, so larger T should resemble numerical scoring more. The experiments show that this holds even with small values of T.

Going beyond our theoretical analysis for $r = 2$ evaluations, we consider the case where students participate in more than two evaluations. Surprisingly, as seen in Figures 1c and 1d, the comparison between numerical scoring and uniform letter grading flips completely with large values of r: numerical scoring becomes *worse* than uniform letter grading (and ULG_T becomes worse than $\mathrm{ULG}_{T'}$ for $T > T'$) when $\alpha_m < \alpha_d$, but *better* when $\alpha_d < \alpha_m$. This shows that the choice of the grading scheme depends not only on the comparison between the strengths of motivational and demotivational effects (α_m vs α_d) but also, crucially, on the number of evaluations r. When $\alpha_m < \alpha_d$, with fewer evaluations (e.g., courses with fewer tests or curricula with fewer semesters), use of numerical scoring may be recommended, whereas with many evaluations (e.g., courses with frequent tests or curricula with many semesters), use of uniform letter grading with fewer letters may be more appropriate.

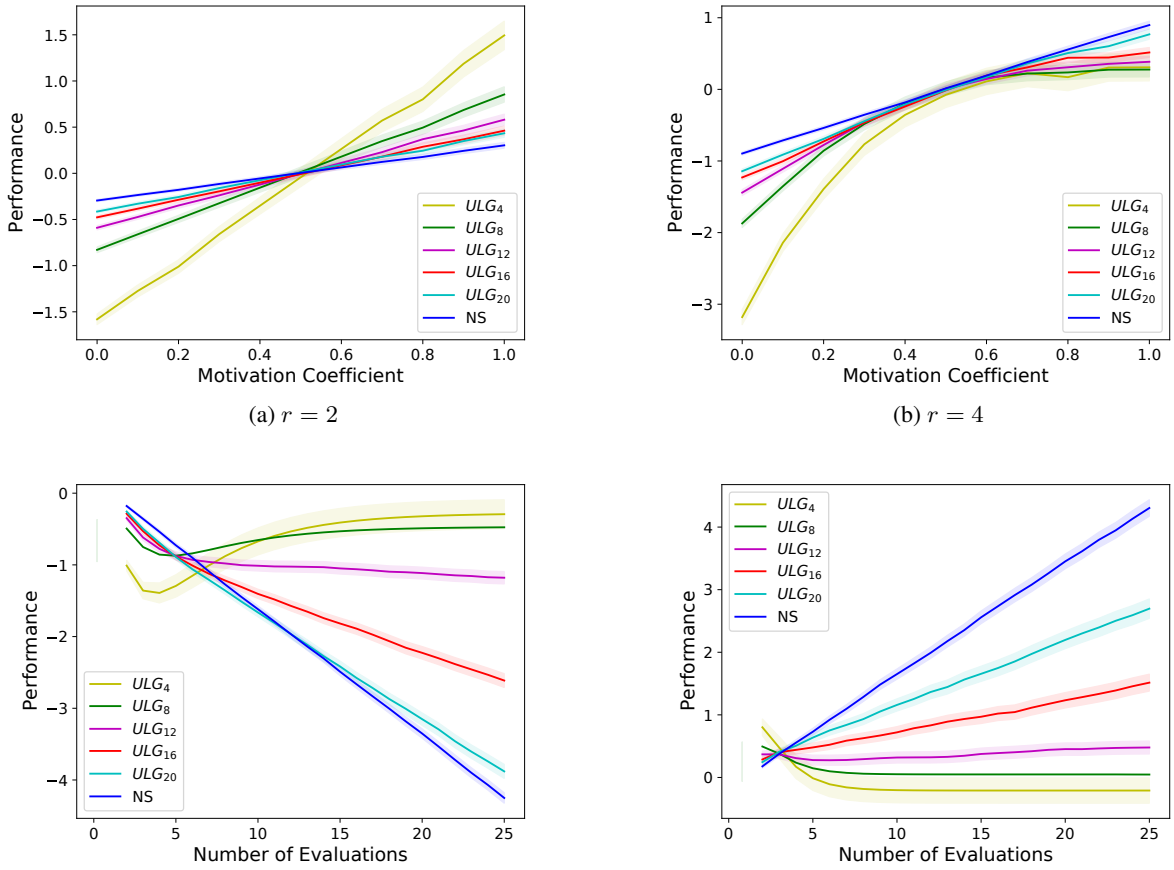
The transition between the regimes of few evaluations and many evaluations is even more surprising. As seen in Figure 1b, with $r = 4$ evaluations, numerical scoring seems to outperform uniform letter grading schemes regardless of the comparison between α_m and α_d. Hence, in general, it is always best to simulate different grading schemes under the model and the number of evaluations of interest in order to pick a suitable grading scheme.

Finally, we observe that under numerical scoring, as the number of evaluations increases, the average student quality declines linearly when $\alpha_m < \alpha_d$ (Figure 1c) and improves linearly when $\alpha_m > \alpha_d$ (Figure 1d). This is expected because it can be shown that under numerical scoring, every evaluation changes the expected student quality by the same amount, which is proportional to $\alpha_m - \alpha_d$, leading to a linear decline or growth. In contrast, under uniform letter grading schemes with very few grades (small T), the average student quality seems to converge and remain stable as the number of evaluations increases, regardless of the comparison between α_m and α_d. This can be explained due to the following stabilizing effect. Let $[\ell, h]$ be a letter grade interval and m be its midpoint. Consider a student who starts with a true quality $q \in [\ell, h]$. The student is likely to receive a score s in the same interval $[\ell, h]$ (so that $(q, s) \in \mathcal{D}$), and thus, a grade of m. This causes the true quality to update in a manner so that it gets closer to m after which the student experiences very little motivation or demotivation due to receiving a grade that is almost equal to her true quality. Of course, the effect is more pronounced when T is small, so letter grade intervals are large compared to the variance of the score model.

Due to the space constraints, we have presented only the most striking observations, deferring the rest to the full version.

5 Discussion

Our work takes the first step towards proposing a statistical model of the impact of letter grading schemes on student per-



(a) $r = 2$ (b) $r = 4$

(c) $\alpha_m = 0.2$ (d) $\alpha_m = 0.8$

Figure 1: Performance of numerical scoring and different uniform letter grading schemes, with $\mu = 65$, $\sigma = 12$, $\gamma = 1.5$ and $\alpha_d = 0.5$ over different motivation coefficients (top) and number of evaluations (bottom). 95% confidence intervals are shown.

formance in sequential evaluations and using it to compare uniform letter grading schemes to numerical scoring. We view our work as a stepping stone and outline 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, they are effectively mapped to the midpoint grade $(\ell + u)/2$. While this is a common method in practice of converting letter grades to percentages [University of Western Ontario, 2022; Victoria University of Wellington, 2022], other values within the range $[\ell, u]$ are also sometimes used [University of Waterloo, 2022].

Non-uniform letter grading. It would be interesting to extend our analysis to non-uniform letter grading schemes. More broadly, how can our model be extended to incorporate truly non-numeric grades (e.g., A, B, etc.) without converting them to numeric grades somehow (e.g., 4, 3.7, etc.)?

Non-linear (de)motivation. 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 study such nonlinear effects.

Exploring implications to pedagogy and beyond. There is a growing literature on optimizing design choices in AI-based learning systems, e.g., algorithmically deciding which explanations to show to students [Zavaleta-Bernuy *et al.*, 2021]. Our insights may inform the design of personalized grading schemes in such systems; they can adjust grade disclosure by learning over time whether students respond more strongly to motivation or demotivation.

More broadly, insights from our work can be explored in other multi-agent systems, such as contest design [Levy *et al.*, 2017] and crowdsourcing [Han *et al.*, 2020], where agents participate in rounds, and feedback from earlier rounds can influence the effort in subsequent rounds. For example, under the right conditions, Theorem 3 may suggest a leaderboard design where teams are grouped into buckets (analogously to letter grading) and their exact performance is not revealed.

References

[Baumeister *et al.*, 2001] Roy F. Baumeister, Ellen Bratslavsky, Catrin Finkenauer, and Kathleen D. Vohs. Bad is stronger than good. *Review of general psychology*, 5(4):323–370, 2001.

[Bloodgood *et al.*, 2020] Robert A. Bloodgood, Jerry G. Short, John M. Jackson, and James R. Martindale. A change to pass/fail grading in the first two years at one medical school results in improved psychological well-being. *Economics of Education Review*, 84:655–662, 2020.

[Brownback, 2018] Andy Brownback. A classroom experiment on effort allocation under relative grading. *Economics of Education Review*, 62:113–128, 2018.

[Cameron and Pierce, 1994] Judy Cameron and W. David Pierce. Reinforcement, reward, and intrinsic motivation: A meta-analysis. *Review of Educational research*, 64(3):363–423, 1994.

[Christensen and Menzel, 1998] Laura J Christensen and Kent E Menzel. The linear relationship between student reports of teacher immediacy behaviors and perceptions of state motivation, and of cognitive, affective, and behavioral learning. *Communication Education*, 47(1), 1998.

[Coleman *et al.*, 1987] Lerita M. Coleman, Lee Jussim, and Jack Abraham. Students' reactions to teachers' evaluations: The unique impact of negative feedback. *Journal of Applied Social Psychology*, 17(12):1051–1070, 1987.

[Czibor *et al.*, 2020] Eszter Czibor, Sander Onderstal, Randolph Sloof, and Mirjam C. van Praag. Does relative grading help male students? Evidence from a field experiment in the classroom. *Economics of Education Review*, 75:101953, 2020.

[Deci *et al.*, 1999] Edward L. Deci, Richard Koestner, and Richard M. Ryan. A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological bulletin*, 125(6):627, 1999.

[Dev, 1997] Poonam C. Dev. Intrinsic motivation and academic achievement: What does their relationship imply for the classroom teacher? *Remedial and special education*, 18(1):12–19, 1997.

[Dubey and Geanakoplos, 2010] Pradeep Dubey and John Geanakoplos. Grading exams: 100, 99, 98,... or A, B, C? *Games and Economic Behavior*, 69(1):72–94, 2010.

[Gray and Bunte, 2022] Thomas Gray and Jonas Bunte. The effect of grades on student performance: Evidence from a quasi-experiment. *College Teaching*, 70(1):15–28, 2022.

[Guskey, 2011] Thomas R Guskey. Five obstacles to grading reform. *Educational Leadership*, 69(3):16, 2011.

[Han *et al.*, 2020] Guangyang Han, Jinzheng Tu, Guoxian Yu, Jun Wang, and Carlotta Domeniconi. Crowdsourcing with multiple-source knowledge transfer. In *Proceedings of the 29th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2908–2914, 2020.

[Kahneman and Tversky, 1979] Daniel Kahneman and Amos Tversky. On the interpretation of intuitive probability: A reply to Jonathan Cohen. *Cognition*, 7(4):409–411, 1979.

[Latham and Locke, 2007] Gary P Latham and Edwin A Locke. New developments in and directions for goal-setting research. *European Psychologist*, 12(4):290, 2007.

[Levy *et al.*, 2017] Priel Levy, David Sarne, and Igor Rochlin. Contest design with uncertain performance and costly participation. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 302–309, 2017.

[Main and Ost, 2014] Joyce B. Main and Ben Ost. The impact of letter grades on student effort, course selection, and major choice: A regression-discontinuity analysis. *The Journal of Economic Education*, 45(1):1–10, 2014.

[McEwan *et al.*, 2021] Patrick J McEwan, Sheridan Rogers, and Akila Weerapana. Grade sensitivity and the economics major at a women's college. In *AEA papers and proceedings*, volume 111, pages 102–06, 2021.

[Nisbet, 1975] John Nisbet. Adding and averaging grades. *Educational Research*, 17(2):95–100, 1975.

[Paredes, 2017] Valentina Paredes. Grading system and student effort. *Education Finance and Policy*, 12(1):107–128, 2017.

[Rohe *et al.*, 2006] Daniel E. Rohe, Patricia A. Barrier, Matthew M. Clark, David A. Cook, Kristin S. Vickers, and Paul A. Decker. The benefits of pass-fail grading on stress, mood, and group cohesion in medical students. *Mayo Clinic Proceedings*, 81(11):1443–1448, 2006.

[Sikora, 2015] Adam S. Sikora. Mathematical theory of student assessment through grading. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.714.8666&rep=rep1&type=pdf, 2015. Accessed: 2022-08-15.

[University of Waterloo, 2022] University of Waterloo. University of Waterloo graduate studies, grading scheme prior to Fall 2001. https://uwaterloo.ca/graduate-studies-academic-calendar/general-information-and-regulations/grades-and-grading, 2022. Accessed: 2022-08-15.

[University of Western Ontario, 2022] University of Western Ontario. University of Western Ontario grading scheme. https://registrar.uwo.ca/academics/grades_progression_and_graduation.html#gpa, 2022. Accessed: 2022-08-15.

[Victoria University of Wellington, 2022] Victoria University of Wellington. Victoria University of Wellington grading scheme. https://www.wgtn.ac.nz/students/study/progress/grades, 2022. Accessed: 2022-08-15.

[Wedell *et al.*, 1989] Douglas H Wedell, Allen Parducci, and Diana Roman. Student perceptions of fair grading: A range-frequency analysis. *The American Journal of Psychology*, pages 233–248, 1989.

[Zavaleta-Bernuy *et al.*, 2021] Angela Zavaleta-Bernuy, Qi Yin Zheng, Hammad Shaikh, Jacob Nogas, Anna Rafferty, Andrew Petersen, and Joseph Jay Williams. Using adaptive experiments to rapidly help students. In *Proceedings of the 22nd International Conference on Artificial Intelligence in Education (AIED)*, pages 422–426, 2021.