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

Evi Micha,[1] **Shreyas Sekar,** [2] **Nisarg Shah** [1]

[1]Department of Computer Science, University of Toronto
[2] Department of Management, University of Toronto
emicha@cs.toronto.edu, shreyas.sekar@rotman.toronto.edu, 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 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 identify other realistic conditions under which uniform letter grading schemes outperform numerical scoring.

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 a specific score or performance level based on their intrinsic ability which directly impacts their performance. We refer to this anchor as the intrinsic quality.
2. **(De)Motivation**: When the student's 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, Koestner, and Ryan 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 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 quality $q_1 = q_2 = 85$. 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 effective intrinsic quality 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 quality 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 would inform the third

[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.

student that their performance was actually close to their intrinsic quality.

This example crucially illustrates that the way that students perceive a non-numerical grade plays a key role, and this often depends on how such grades are perceived in the outside world. We study a scheme of mapping letter grades to percentages, which is widely used in practice (see, e.g., (University of Western Ontario 2022; Victoria University of Wellington 2022)), where the grade of the student is given by the midpoint of the interval containing all the scores that map to that letter grade. For example, if a student receives $A-$, which captures all scores in the range $[80, 90]$, the student effectively receives a grade of 85, and this grade is what the student compares to her intrinsic quality.

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). Note that under uniform (or even non-uniform) letter grading schemes, one cannot simply assign all students a grade of 100 to maximally motivate them: for example, if the entire range of $[0, 100]$ is mapped to one letter grade ($T = 1$), all students would receive a grade of 50 due to the midpoint representation.

First, we theoretically study the case where two sequential evaluations take place. 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. By assuming additional conditions, such as when the intrinsic qualities of the students follow a uniform distribution, we can conclude that 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, Jussim, and Abraham 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 that 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.

Related work. There are 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 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. Future work may combine the two approaches by modeling student motivation as a function of both the grading scheme to be used in the present evaluation as well as performance in the past evaluations.

Another related work is that of Sikora (2015), who also compares grading schemes, but his goal is to study the tradeoff 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 our use of the midpoint representation; e.g., as mentioned before, it prevents the instructor from simply assigning a grade of 100 to all the students regardless of their scores.

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.

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 $B : [0, 1] \to [0, 1]$ that maps the score to a grade.

Letter grading. A *letter grading* scheme $B_{\vec{c}}$ is specified by a vector $\vec{c} = (c_0 = 0, c_1, \ldots, c_{T-1}, c_T = 1)$, 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 $s \in [0, 1)$,[2]

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

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.

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.[3]

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,

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

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.

Uniform Letter Grading vs Numerical Scoring

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. We 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 necessarily 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.

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

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

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 (but with different coefficients). Hence,

$$\begin{aligned}&\mathbb{E}\Big[h(q, B(s)) + h(1-q, B(1-s))\Big]\\ &\quad= (\alpha_m - \alpha_d) \cdot \mathbb{E}\Big[|q - s|\Big].\end{aligned}$$

Plugging this into Equation (2), we obtain the desired 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 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 $\mathbb{E}[|q - \mathrm{NS}(s)| - |q - \mathrm{ULG}_T(s)|]$, 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 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. 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 $\mathbb{E}\Big[|q - s|\Big] \leqslant \mathbb{E}\Big[|q - \mathrm{ULG}_T(s)|\Big]$. 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.

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. Fix any $q \in [0, 1]$. Note that

$$\begin{aligned}&\mathbb{E}\Big[|q - s| \;\Big|\; (q, s) \notin \mathcal{D}\Big]\\ &\quad= \Pr\Big[\mathrm{ULG}_T(s) < \mathrm{ULG}_T(q) \;\Big|\; (q, s) \notin \mathcal{D}\Big]\\ &\qquad\quad\cdot \mathbb{E}\Big[q - s \;\Big|\; \mathrm{ULG}_T(s) < \mathrm{ULG}_T(q)\Big]\\ &\qquad+ \Pr\Big[\mathrm{ULG}_T(s) > \mathrm{ULG}_T(q) \;\Big|\; (q, s) \notin \mathcal{D}\Big]\\ &\qquad\quad\cdot \mathbb{E}\Big[s - q \;\Big|\; \mathrm{ULG}_T(s) > \mathrm{ULG}_T(q)\Big]\\ &\leqslant \Pr\Big[\mathrm{ULG}_T(s) < \mathrm{ULG}_T(q) \;\Big|\; (q, s) \notin \mathcal{D}\Big]\\ &\qquad\quad\cdot \mathbb{E}\Big[q - \mathrm{ULG}_T(s) \;\Big|\; \mathrm{ULG}_T(s) < \mathrm{ULG}_T(q)\Big]\\ &\qquad+ \Pr\Big[\mathrm{ULG}_T(s) > \mathrm{ULG}_T(q) \;\Big|\; (q, s) \notin \mathcal{D}\Big]\\ &\qquad\quad\cdot \mathbb{E}\Big[\mathrm{ULG}_T(s) - q \;\Big|\; \mathrm{ULG}_T(s) > \mathrm{ULG}_T(q)\Big]\\ &\quad= \mathbb{E}\Big[|q - \mathrm{ULG}_T(s)| \;\Big|\; (q, s) \notin \mathcal{D}\Big],\end{aligned}$$

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

$$\mathbb{E}\Big[s \;\Big|\; \mathrm{ULG}_T(s) < \mathrm{ULG}_T(q)\Big] \geqslant \mathbb{E}\Big[\mathrm{ULG}_T(s) \;\Big|\; \mathrm{ULG}_T(s) < \mathrm{ULG}_T(q)\Big],$$

and

$$\mathbb{E}\Big[s \;\Big|\; \mathrm{ULG}_T(s) > \mathrm{ULG}_T(q)\Big] \leqslant \mathbb{E}\Big[\mathrm{ULG}_T(s) \;\Big|\; \mathrm{ULG}_T(s) > \mathrm{ULG}_T(q)\Big].$$

This completes the proof. □

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, given in the next definitions, jointly with the score model $\mathcal{S}$, 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 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 $\mathcal{S}$ 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 $\Pr[(q,s) \in \mathcal{D}^{\mathrm{opp}}]$. For further intuition regarding the comparison between $\Pr[(q,s) \in \mathcal{D}^{\mathrm{same}}]$ and $\Pr[(q,s) \in \mathcal{D}^{\mathrm{opp}}]$, see Figure 2 in Appendix A.

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 needed. However, using very different techniques, we can show that even a constant of 3 suffices to obtain the same result under strong symmetry. This broadens the scope of the result to include slightly 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 $\mathcal{S}$.*

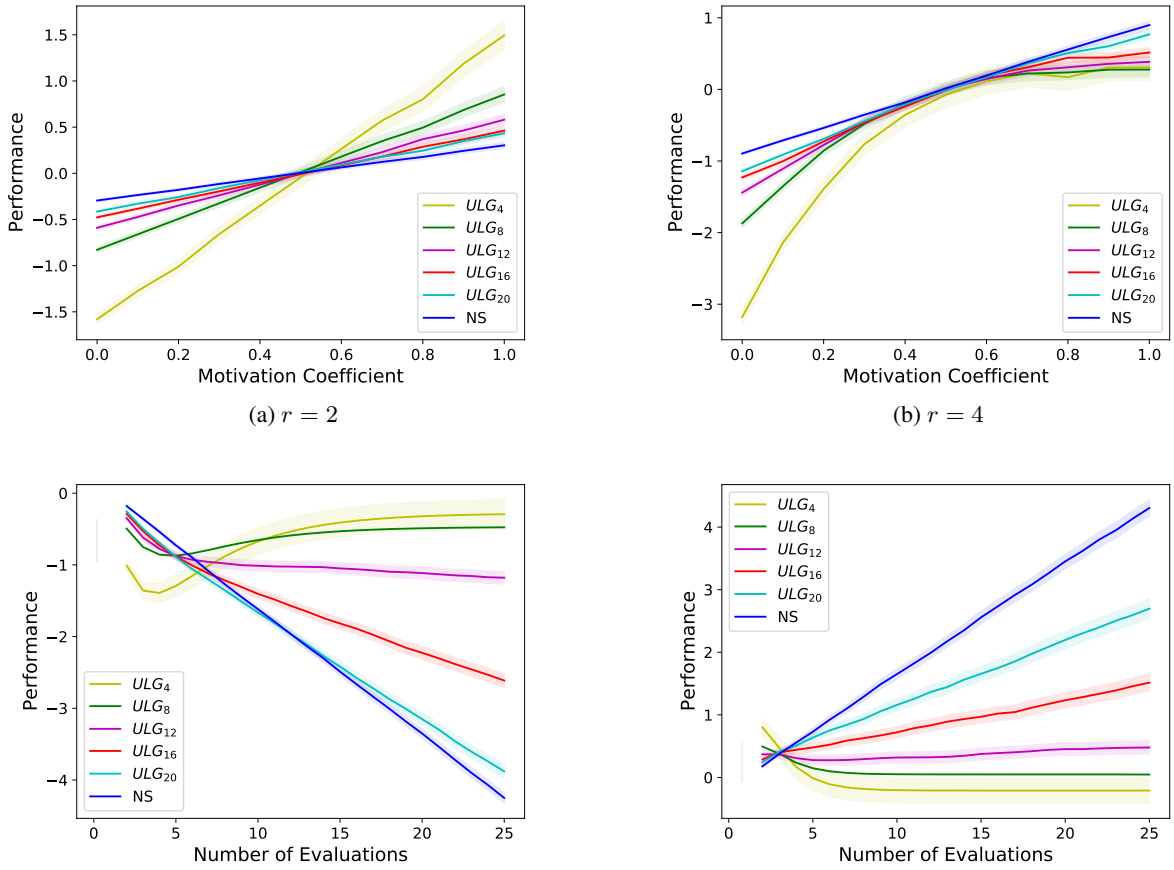
1. *$\mathcal{Q}$ is uniform over $[0,1]$;*
2. *$\mathcal{S}$ is strongly symmetric; and*
3. $\Pr\Big[(q,s) \in \mathcal{D}^{\mathrm{same}}\Big] \geqslant 3 \cdot \Pr\Big[(q,s) \in \mathcal{D}^{\mathrm{opp}}\Big]$.

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

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.

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.



(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.

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, Parducci, and Roman 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 very-well concentrated around the true quality. Finally, our theoretical analysis was limited to two sequential 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:

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

We remark that measuring average improvement in quality, $\mathbb{E}[(1/r)\sum_{j=1}^{r} q_j - q_1]$, yield qualitatively the same results, so we omit them from the paper.

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 a more realistic distribution than the uniform distribution assumed in Theorems 3 and 4. 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. 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 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, leading it to converge to m over many evaluations. Once the true quality becomes very close to m, the student experiences very little motivation or demotivation due to receiving a grade that is almost equal to her true quality with high probability. 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 empirical observations here; the rest are deferred to Appendix C, where we notice that the first implication of Corollary 1 continues to hold even when the score is not well-concentrated around the true quality.

Discussion

Our work takes the first step towards proposing a statistical model of the psychological impact of letter grading schemes on student performance in sequential evaluations and using it to compare uniform letter grading schemes to numerical scoring. Obviously, we model one specific (de)motivational effect and future work must combine it with many other well-known effects in educational settings, but we view our work as a stepping stone 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, they are effectively mapped to the midpoint grade $(\ell + u)/2$. This is a common method in practice of converting letter grades to percentages (University of Western Ontario 2022; Victoria University of Wellington 2022), but other values within the range $[\ell, u]$ are also sometimes used (University of Waterloo 2022).

Non-uniform letter grading. Our analysis is limited to *uniform* letter grading schemes, which are used less often in practice. It would be interesting to extend our analysis to non-uniform letter 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 converting them to numeric grades somehow (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 nonlinear (de)motivational effects.

References

Baumeister, R. F.; Bratslavsky, E.; Finkenauer, C.; and Vohs, K. D. 2001. Bad is stronger than good. *Review of general psychology*, 5(4): 323–370.

Bloodgood, R. A.; Short, J. G.; Jackson, J. M.; and Martindale, J. R. 2020. 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.

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

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

Coleman, L. M.; Jussim, L.; and Abraham, J. 1987. Students' R] eactions to Teachers' E] valuations: The Unique Impact of Negative Feedback 1. *Journal of Applied Social Psychology*, 17(12): 1051–1070.

Czibor, E.; Onderstal, S.; Sloof, R.; and van Praag, M. C. 2020. Does relative grading help male students? Evidence from a field experiment in the classroom. *Economics of Education Review*, 75: 101953.

Deci, E. L.; Koestner, R.; and Ryan, R. M. 1999. A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological bulletin*, 125(6): 627.

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

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

Gray, T.; and Bunte, J. 2022. The Effect of Grades on Student Performance: Evidence from a Quasi-Experiment. *College Teaching*, 70(1): 15–28.

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

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

Main, J. B.; and Ost, B. 2014. 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.

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

Rohe, D.; Barrier, P.; Clark, M.; Cook, D.; Vickers, K.; and Decker, P. A. 2006. The benefits of pass-fail grading on stress, mood, and group cohesion in medical students. *Mayo Clinic Proceedings*, 81(11): 1443–1448.

Sikora, A. S. 2015. Mathematical Theory of Student Assessment Through Grading. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.714.8666&rep=rep1&type=pdf. Accessed: 2022-08-15.

University of Waterloo. 2022. 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. Accessed: 2022-08-15.

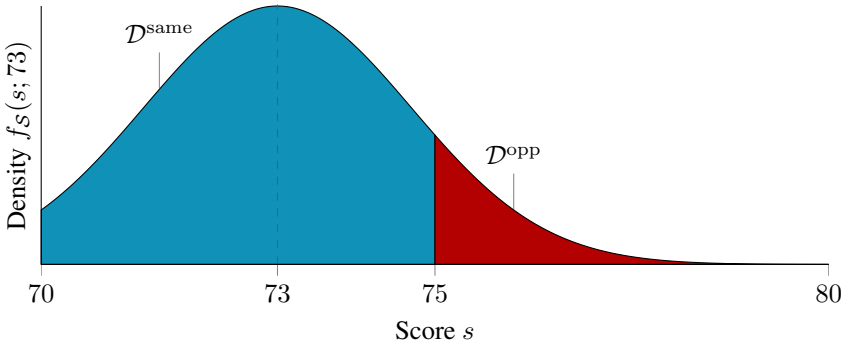
University of Western Ontario. 2022. University of Western Ontario Grading Scheme. https://registrar.uwo.ca/academics/grades_progression_and_graduation.html#gpa. Accessed: 2022-08-15.

Victoria University of Wellington. 2022. Victoria University of Wellington Grading Scheme. https://www.wgtn.ac.nz/students/study/progress/grades. Accessed: 2022-08-15.

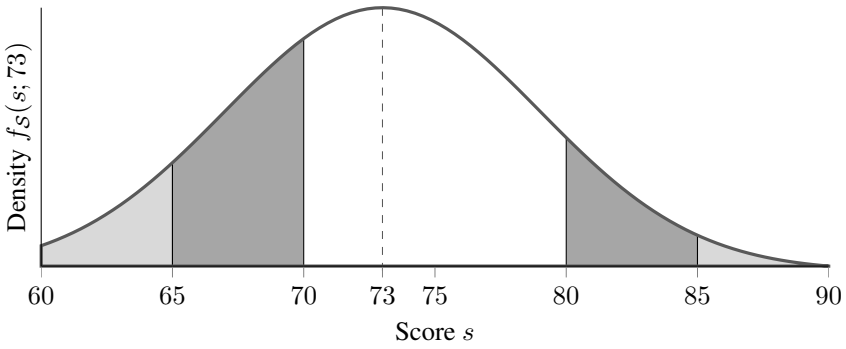
Wedell, D. H.; Parducci, A.; and Roman, D. 1989. Student perceptions of fair grading: A range-frequency analysis. *The American Journal of Psychology*, 233–248.

Appendix

A Intuition Regarding $\mathcal{D}^{\text{same}}$ vs $\mathcal{D}^{\text{opp}}$ & Single-Peakedness



(a) With a small value of γ, 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 $\mathcal{D}^{\text{same}}$ while the latter contributes to $\mathcal{D}^{\text{opp}}$. 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 $\Pr[(q, s) \in \mathcal{D}^{\text{same}}]$ to be sufficiently higher than $\Pr[(q, s) \in \mathcal{D}^{\text{opp}}]$, 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 2: Both figures show the probability density function of the score distribution $\mathcal{S}(q)$ 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 $\Pr[(q, s) \in \mathcal{D}^{\text{same}}]$ to be sufficiently higher than $\Pr[(q, s) \in \mathcal{D}^{\text{opp}}]$. The bottom figure conveys the intuition behind the observation used at the end of the proof of Theorem 2.

B Missing Proofs

Useful Lemmas

Before we dive into the missing proofs, we state the integral version of the well-known Chebyshev's inequality and its two useful implications.

Lemma 1 (Integral Chebyshev Inequality)**.** *If functions $f, g : [a, b] \to \mathbb{R}_{\geqslant 0}$ are either both non-increasing or both non-decreasing, then*

$$\frac{1}{b-a}\int_a^b f(x)g(x)\,\mathrm{d}x \geqslant \left(\frac{1}{b-a}\int_a^b f(x)\,\mathrm{d}x\right)\cdot\left(\frac{1}{b-a}\int_a^b g(x)\,\mathrm{d}x\right).$$

If one of them is non-decreasing while the other is non-increasing, the inequality is reversed.

The following inequality is obtained by substituting $f(x) = x$ (and thus, $\frac{1}{b-a}\int_a^b f(x)\,\mathrm{d}x = \frac{a+b}{2}$) into Lemma 1.

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

$$\int_a^b xg(x)\,\mathrm{d}x \leqslant \frac{a+b}{2} \cdot \int_a^b g(x)\,\mathrm{d}x,$$

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

If g is a probability density function over $[a, b]$, then $\int_a^b g(x)\,\mathrm{d}x = 1$, yielding the following (quite natural) implication.

Lemma 3. *Let X be a random variable over $[a, b]$ with a non-increasing probability density function $g : [a, b] \to \mathbb{R}_{\geqslant 0}$. Then, $\mathbb{E}[X] \leqslant (a + b)/2$, 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 $g : [a, b] \to \mathbb{R}_{\geqslant 0}$ be a concave function with $g(b) = 0$. Then, we have*

$$\int_a^b (b-x)g(x)\,\mathrm{d}x \leqslant \frac{2(b-a)}{3}\int_a^b g(x)\,\mathrm{d}x.$$

Proof. Due to concavity of g, we have

$$\int_x^b g(t)\,\mathrm{d}t \geqslant \frac{1}{2}(b-x)g(x).$$

Hence, we have

$$\begin{aligned}
\int_a^b \frac{1}{2}(b-x)g(x)\,\mathrm{d}x &\geqslant \int_{x=a}^b \int_{t=x}^b g(t)\,\mathrm{d}t\,\mathrm{d}x \\
&= \int_{t=a}^b \int_{x=a}^t g(t)\,\mathrm{d}x\,\mathrm{d}t && \text{(Fubini's theorem)} \\
&= \int_{t=a}^b (t-a)g(t)\,\mathrm{d}t \\
&= \int_{x=a}^b (x-a)g(x)\,\mathrm{d}x && \text{(Change of variable name)} \\
&= \int_a^b (b-a)g(x)\,\mathrm{d}x - \int_a^b (b-x)g(x)\,\mathrm{d}x.
\end{aligned}$$

Rearranging the terms yields the desired inequality. □

Proof of Thereom 3

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

$$\begin{aligned}
&\mathbb{E}\Big[|q-s| - |q - \mathrm{ULG}_T(s)| \;\Big|\; (q,s) \in \mathcal{D}\Big] \\
&\quad = \Pr[(q,s) \in \mathcal{D}^{\mathrm{same}} \;\Big|\; (q,s) \in \mathcal{D}] \cdot \mathbb{E}\Big[|q-s| - |q - \mathrm{ULG}_T(s)| \;\Big|\; (q,s) \in \mathcal{D}^{\mathrm{same}}\Big] \\
&\qquad + \Pr[(q,s) \in \mathcal{D}^{\mathrm{opp}} \;\Big|\; (q,s) \in \mathcal{D}] \cdot \mathbb{E}\Big[|q-s| - |q - \mathrm{ULG}_T(s)| \;\Big|\; (q,s) \in \mathcal{D}^{\mathrm{opp}}\Big] \\
&\quad \leqslant 0.
\end{aligned} \tag{3}$$

Let us analyze the expected value of $|q-s| - |q - \mathrm{ULG}_T(s)|$ conditioned on both $(q,s) \in \mathcal{D}^{\mathrm{same}}$ and $(q,s) \in \mathcal{D}^{\mathrm{opp}}$ separately.

Analyzing $\mathcal{D}^{\mathrm{same}}$. 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

$$\begin{aligned}
\mathcal{D}^{\mathrm{same}} = \Big\{(q,s) :&(\ell(k) \leqslant q \leqslant s \leqslant m(k)) \vee (\ell(k) \leqslant s \leqslant q \leqslant m(k)) \vee \\
&(m(k) \leqslant q \leqslant s < h(k)) \vee (m(k) \leqslant s \leqslant q < h(k)),\ k \in \{0, 1, \ldots, T-1\}\Big\}.
\end{aligned}$$

Fix an arbitrary $k \in \{0, 1, \ldots, T-1\}$; write ℓ, m, and h while omitting the fixed k in the argument; and let us analyze the desired expression $|q-s| - |q - \mathrm{ULG}_T(s)|$ 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 $(q,s) \in \mathcal{D}^{\text{same}}$ (i.e., aggregated across all k). Note that in each case, we have $\mathrm{ULG}_T(q) = \mathrm{ULG}_T(s) = m$.

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

$$\begin{aligned}
\mathbb{E}\Big[s - m \mid \ell \leqslant q \leqslant s \leqslant m\Big] &= \frac{\int_{q=\ell}^{m}\int_{s=q}^{m} f_{\mathcal{Q}}(q)\cdot f_{\mathcal{S}}(s;q)\cdot(s-m)\,\mathrm{d}s\,\mathrm{d}q}{\Pr[\ell \leqslant q \leqslant s \leqslant m]} \\
&= \frac{\int_{q=\ell}^{m} 1 \cdot \mathbb{E}\Big[s-m \;\Big|\; q, s\in[q,m]\Big]\cdot \Pr\Big[s\in[q,m]\;\Big|\; q\Big]\,\mathrm{d}q}{\int_{q=\ell}^{m} 1\cdot \Pr\Big[s\in[q,m]\;\Big|\; q\Big]\,\mathrm{d}q} \\
&\leqslant \frac{-\int_{q=\ell}^{m}\left(\frac{m-q}{2}\right)\cdot \Pr\Big[s\in[q,m]\;\Big|\; q\Big]\,\mathrm{d}q}{\int_{q=\ell}^{m} 1\cdot \Pr\Big[s\in[q,m]\;\Big|\; q\Big]\,\mathrm{d}q} \\
&\leqslant \frac{-\frac{1}{m-\ell}\cdot\left(\int_{q=\ell}^{m}\frac{m-q}{2}\,\mathrm{d}q\right)\cdot\left(\int_{q=\ell}^{m}\Pr\Big[s\in[q,m]\;\Big|\; q\Big]\,\mathrm{d}q\right)}{\int_{q=\ell}^{m} 1\cdot \Pr\Big[s\in[q,m]\;\Big|\; q\Big]\,\mathrm{d}q} \\
&= -\frac{1}{(\Delta/2)}\int_{r=0}^{\frac{\Delta}{2}}\frac{r}{2}\,\mathrm{d}r = -\frac{\Delta}{8}.
\end{aligned}$$

   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 3). Hence, $\mathbb{E}[s|q, s \in [q,m]] \leqslant (q+m)/2$. The fourth transition is the integral Chebyshev inequality (Lemma 1), which holds because both $(m-q)/2$ and $\Pr\Big[s\in[q,m]\;\Big|\; q\Big]$ are non-negative, non-increasing functions of q in $[\ell, m]$ (Assumption 3).
2. $\ell \leqslant s \leqslant q \leqslant m$. In this case, $|q-s| - |q - \mathrm{ULG}_T(s)| = 2q - m - s$. Note that

$$\begin{aligned}
&\mathbb{E}\Big[2q - m - s \mid \ell \leqslant s \leqslant q \leqslant m\Big] \\
&\quad= \frac{\int_{q=\ell}^{m}\int_{s=\ell}^{q} f_{\mathcal{Q}\times\mathcal{S}}(q,s)\cdot(2q-m-s)\,\mathrm{d}s\,\mathrm{d}q}{\Pr[\ell \leqslant s \leqslant q \leqslant m]} \\
&\quad= \frac{\int_{s=\ell}^{m} f_{\mathcal{S}}(s)\mathbb{E}\Big[2q-m-s \;\Big|\; s, q\in[s,m]\Big]\cdot \Pr\Big[q\in[s,m]\;\Big|\; s\Big]\,\mathrm{d}s}{\Pr[\ell \leqslant s \leqslant q \leqslant m]}.
\end{aligned}$$

   Here, we use $f_{\mathcal{Q}\times\mathcal{S}}(q,s) = f_{\mathcal{Q}}(q)\cdot f_{\mathcal{S}}(s;q)$ to denote the joint probability density of q and s, and $f_{\mathcal{S}}(s) = \int_{q=0}^{1} f_{\mathcal{Q}}(q) f_{\mathcal{S}}(s;q)\,\mathrm{d}q$ to denote the marginal probability density of s.

   We argue that $\mathbb{E}\Big[2q-m-s \;\Big|\; s, q\in[s,m]\Big] \leqslant 0$. Intuitively, this is because the posterior distribution of $q \in [s,m]$ conditioned on a fixed value of s and on $q \in [s,m]$ is single-peaked with peak at s by Assumptions 1 and 3. Hence, $\mathbb{E}\Big[q \;\Big|\; s, q\in[s,m]\Big] \leqslant (s+m)/2$. Formally, this can be viewed as

$$\begin{aligned}
\mathbb{E}\Big[2q-m-s \;\Big|\; s, q\in[s,m]\Big] &= \frac{\int_{q=s}^{m} f(q;s)(2q-m-s)\,\mathrm{d}q}{\int_{q=s}^{m} f(q;s)\,\mathrm{d}q} \\
&\leqslant \frac{\frac{1}{m-s}\cdot\left(\int_{q=s}^{m} f(q;s)\,\mathrm{d}q\right)\cdot\left(\int_{q=s}^{m}(2q-m-s)\,\mathrm{d}q\right)}{\int_{q=s}^{m} f(q;s)\,\mathrm{d}q} = 0,
\end{aligned}$$

   where $f(q;s) = \frac{f_{\mathcal{Q}}(q)\cdot f_{\mathcal{S}}(s;q)}{f_{\mathcal{S}}(s)}$ 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;[4,5] and the final transition holds because the second integral in the numerator is 0.

[4] To see why $f(q;s) = \frac{f_{\mathcal{Q}}(q)\cdot f_{\mathcal{S}}(s;q)}{f_{\mathcal{S}}(s)}$ 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).

[5] Technically, integral Chebyshev inequality requires non-negative functions, and $2q-(m+s)$ can be negative when $q < (m+s)/2$.

3. $m \leqslant q \leqslant s < h$. In this case, $|q-s| - |q - \mathrm{ULG}_T(s)| = m + s - 2q$. Due to the same reasoning as in Case 2, we have that $\mathbb{E}\Big[m + s - 2q \;\Big|\; m \leqslant q \leqslant s < h\Big] \leqslant 0$.
4. $m \leqslant s \leqslant q < h$. In this case, $|q-s| - |q - \mathrm{ULG}_T(s)| = m - s$. Due to the same reasoning as in Case 1, we have that $\mathbb{E}\Big[m - s \;\Big|\; m \leqslant s \leqslant q < h\Big] \leqslant -\Delta/8$.

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 $f_{\mathcal{S}}(a; b)/f_{\mathcal{S}}(b; a) \leqslant \gamma$ for all $a, b \in [0, 1]$, it follows that $p_1 \geqslant p_2/\gamma$ and $p_4 \geqslant p_3/\gamma$. Hence, $p_1 + p_4 \geqslant (p_2 + p_3)/\gamma$. Using $p_1 + p_2 + p_3 + p_4 = 1$, we get $p_1 + p_4 \geqslant 1/(\gamma + 1)$.

Combining the analysis from the four cases above, we have

$$\mathbb{E}\Big[|q-s| - |q - \mathrm{ULG}_T(s)| \;\Big|\; (q, s) \in \mathcal{D}^{\text{same}}\Big] \leqslant -(p_1 + p_4) \cdot \frac{\Delta}{8} \leqslant -\frac{\Delta}{8(\gamma + 1)}. \tag{4}$$

Analyzing $\mathcal{D}^{\text{opp}}$. Note that

$$\mathcal{D}^{\text{opp}} = \cup_{k \in \{0,1,\ldots,T-1\}} \left\{(q, s) : (\ell(k) \leqslant q \leqslant m(k) \leqslant s \leqslant h(k)) \vee (\ell(k) \leqslant s \leqslant m(k) \leqslant q \leqslant h(k))\right\}.$$

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

1. $\ell \leqslant q \leqslant m \leqslant s \leqslant h$: In this case, we have $|q-s| - |q - \mathrm{ULG}_T(s)| = s - m$. Note that
$$\mathbb{E}\Big[s - m \;\Big|\; \ell \leqslant q \leqslant m \leqslant s \leqslant h\Big] \leqslant \Delta/4. \tag{5}$$
This is because $s \in [m, m + \Delta/2]$ and, due to single-peakedness of the score model and $q \leqslant m$, it is at most $m + \Delta/4$ in expectation.
2. $\ell \leqslant s \leqslant m \leqslant q \leqslant h$: In this case, we have $|q-s| - |q - \mathrm{ULG}_T(s)| = m - s$, and the same reasoning as above shows that
$$\mathbb{E}\Big[m - s \;\Big|\; \ell \leqslant s \leqslant m \leqslant q \leqslant h\Big] \leqslant \Delta/4. \tag{6}$$

Combining Equations (5) and (6) and aggregating over all $k \in \{0, 1, \ldots, T-1\}$, we get that

$$\mathbb{E}\Big[|q-s| - |q - \mathrm{ULG}_T(s)| \;\Big|\; (q, s) \in \mathcal{D}^{\text{opp}}\Big] \leqslant \Delta/4. \tag{7}$$

Finally, combining Equations (4) and (7), we have that

$$\begin{aligned}
&\mathbb{E}\Big[|q-s| - |q - \mathrm{ULG}_T(s)| \;\Big|\; (q, s) \in \mathcal{D}\Big] \\
&\quad\leqslant \Pr\Big[(q, s) \in \mathcal{D}^{\text{same}} \;\Big|\; (q, s) \in \mathcal{D}\Big] \cdot \left(-\frac{\Delta}{8(\gamma + 1)}\right) + \Pr\Big[(q, s) \in \mathcal{D}^{\text{opp}} \;\Big|\; (q, s) \in \mathcal{D}\Big] \cdot \frac{\Delta}{4} \leqslant 0,
\end{aligned}$$

where the final transition holds because $\Pr\Big[(q, s) \in \mathcal{D}^{\text{same}}\Big] \geqslant 2(\gamma + 1) \cdot \Pr\Big[(q, s) \in \mathcal{D}^{\text{opp}}\Big]$ (Assumption 4), which is equivalent to

$$\Pr\Big[(q, s) \in \mathcal{D}^{\text{same}} \;\Big|\; (q, s) \in \mathcal{D}\Big] \geqslant 2(\gamma + 1) \cdot \Pr\Big[(q, s) \in \mathcal{D}^{\text{same}} \;\Big|\; (q, s) \in \mathcal{D}\Big].$$

This completes the proof. □

Proof of Theorem 4

Proof. As in the proof of Theorem 3, note that given Theorem 2, we only need to prove

$$\begin{aligned}
&\mathbb{E}\Big[|q-s| - |q - \mathrm{ULG}_T(s)| \;\Big|\; (q, s) \in \mathcal{D}\Big] \\
&\quad= \Pr[(q, s) \in \mathcal{D}^{\text{same}} \;\Big|\; (q, s) \in \mathcal{D}] \cdot \mathbb{E}\Big[|q-s| - |q - \mathrm{ULG}_T(s)| \;\Big|\; (q, s) \in \mathcal{D}^{\text{same}}\Big] \\
&\qquad+ \Pr[(q, s) \in \mathcal{D}^{\text{opp}} \;\Big|\; (q, s) \in \mathcal{D}] \cdot \mathbb{E}\Big[|q-s| - |q - \mathrm{ULG}_T(s)| \;\Big|\; (q, s) \in \mathcal{D}^{\text{opp}}\Big] \\
&\quad\leqslant 0.
\end{aligned} \tag{8}$$

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.

In the proof of Theorem 3, we analyzed the expected value of $|q-s|-|q-\mathrm{ULG}_T(s)|$ conditioned on both $(q,s)\in\mathcal{D}^{\mathrm{same}}$ and $(q,s)\in\mathcal{D}^{\mathrm{opp}}$ 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 (8) when $\Pr[(q,s)\in\mathcal{D}^{\mathrm{same}}]\geqslant 2(\gamma+1)\cdot\Pr[(q,s)\in\mathcal{D}^{\mathrm{opp}}]$.

With strong symmetry (Assumption 2), we improve the former upper bound to $-\frac{\Delta}{12}$, which improves the sufficient condition to $\Pr[(q,s)\in\mathcal{D}^{\mathrm{same}}]\geqslant 3\cdot\Pr[(q,s)\in\mathcal{D}^{\mathrm{opp}}]$. That is, our goal is to prove

$$\mathbb{E}\Big[|q-s|-|q-\mathrm{ULG}_T(s)|\;\Big|\;(q,s)\in\mathcal{D}^{\mathrm{same}}\Big]\leqslant-\frac{\Delta}{12}.$$

Note that $\mathcal{D}^{\mathrm{same}}=\cup_{k\in\{0,1,\ldots,T-1\}}\mathcal{D}^{\mathrm{same}}_k$, where $\mathcal{D}^{\mathrm{same}}_k=\mathcal{D}^{\mathrm{same}}\cap[k\Delta,(k+1)\Delta)^2$. We show that $\mathbb{E}\Big[|q-s|-|q-\mathrm{ULG}_T(s)|\;\Big|\;(q,s)\in\mathcal{D}^{\mathrm{same}}_k\Big]\leqslant-\frac{\Delta}{12}$ 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 $\mathcal{D}^{\mathrm{same}}_k$ as $\mathcal{D}^{\mathrm{same}}_{k,\mathrm{low}}\cup\mathcal{D}^{\mathrm{same}}_{k,\mathrm{high}}$, where $\mathcal{D}^{\mathrm{same}}_{k,\mathrm{low}}=\{(q,s):\ell\leqslant q,s<m\}$ (both the true quality and the score are lower than the midpoint) and $\mathcal{D}^{\mathrm{same}}_{k,\mathrm{high}}=\{(q,s):m\leqslant q,s<h\}$ (both the true quality and the score are at least as high as the midpoint). Crucially, we note that

$$\mathbb{E}\Big[|q-s|-|q-m|\;\Big|\;(q,s)\in\mathcal{D}^{\mathrm{same}}_{k,\mathrm{low}}\Big]=\mathbb{E}\Big[|q-s|-|q-m|\;\Big|\;(q,s)\in\mathcal{D}^{\mathrm{same}}_{k,\mathrm{high}}\Big].$$

This is because the transformation $(q,s)\to(q',s')$, where $q'=m+(m-q)$ and $s'=m+(m-s)$, is a bijection mapping each point $(q,s)\in\mathcal{D}^{\mathrm{same}}_{k,\mathrm{low}}$ to a point $(q',s')\in\mathcal{D}^{\mathrm{same}}_{k,\mathrm{high}}$ with $|q-s|-|q-m|=|q'-s'|-|q'-m|$ as well as $f_{\mathcal{Q}\times\mathcal{S}}(q,s)=f_{\mathcal{Q}\times\mathcal{S}}(q',s')$; the last observation relies on $\mathcal{Q}$ being a uniform distribution (Assumption 1) and $\mathcal{S}$ being strongly symmetric (Assumption 2).

Hence, it is sufficient to show that

$$\mathbb{E}\Big[|q-s|-|q-m|\;\Big|\;(q,s)\in\mathcal{D}^{\mathrm{same}}_{k,\mathrm{low}}\Big]\leqslant-\frac{\Delta}{12}.$$

Next, we further partition $\mathcal{D}^{\mathrm{same}}_{k,\mathrm{low}}$ as $\mathcal{D}^{\mathrm{same}}_{k,\mathrm{low,inc}}\cup\mathcal{D}^{\mathrm{same}}_{k,\mathrm{low,dec}}$, where $\mathcal{D}^{\mathrm{same}}_{k,\mathrm{low,inc}}=\{(q,s):\ell\leqslant q\leqslant s<m\}$ (the score is at least as much as the true quality) and $\mathcal{D}^{\mathrm{same}}_{k,\mathrm{low,dec}}=\{(q,s):\ell\leqslant s\leqslant q<m\}$ (the score is at most as much as the true quality). Note that

$$\begin{aligned}
&\mathbb{E}\Big[|q-s|-|q-m|\;\Big|\;(q,s)\in\mathcal{D}^{\mathrm{same}}_{k,\mathrm{low}}\Big]\\
&\quad=\Pr\Big[(q,s)\in\mathcal{D}^{\mathrm{same}}_{k,\mathrm{low,inc}}\;\Big|\;(q,s)\in\mathcal{D}^{\mathrm{same}}_{k,\mathrm{low}}\Big]\cdot\mathbb{E}\Big[|q-s|-|q-m|\;\Big|\;(q,s)\in\mathcal{D}^{\mathrm{same}}_{k,\mathrm{low,inc}}\Big]\\
&\quad+\Pr\Big[(q,s)\in\mathcal{D}^{\mathrm{same}}_{k,\mathrm{low,dec}}\;\Big|\;(q,s)\in\mathcal{D}^{\mathrm{same}}_{k,\mathrm{low}}\Big]\cdot\mathbb{E}\Big[|q-s|-|q-m|\;\Big|\;(q,s)\in\mathcal{D}^{\mathrm{same}}_{k,\mathrm{low,dec}}\Big].
\end{aligned}$$

First, we argue that

$$\Pr\Big[(q,s)\in\mathcal{D}^{\mathrm{same}}_{k,\mathrm{low,inc}}\;\Big|\;(q,s)\in\mathcal{D}^{\mathrm{same}}_{k,\mathrm{low}}\Big]=\Pr\Big[(q,s)\in\mathcal{D}^{\mathrm{same}}_{k,\mathrm{low,dec}}\;\Big|\;(q,s)\in\mathcal{D}^{\mathrm{same}}_{k,\mathrm{low}}\Big]=\frac{1}{2}.$$

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

Next, recall that in the proof of Theorem 3 (Case 2 in the analysis of $\mathcal{D}^{\mathrm{same}}$), we had already argued

$$\mathbb{E}\Big[|q-s|-|q-m|\;\Big|\;(q,s)\in\mathcal{D}^{\mathrm{same}}_{k,\mathrm{low,dec}}\Big]=\mathbb{E}[2q-s-m|\ell\leqslant s\leqslant q<m]\leqslant 0.$$

Hence, we have

$$\mathbb{E}\Big[|q-s|-|q-m|\;\Big|\;(q,s)\in\mathcal{D}^{\mathrm{same}}_{k,\mathrm{low}}\Big]\leqslant\frac{1}{2}\cdot\mathbb{E}\Big[|q-s|-|q-m|\;\Big|\;(q,s)\in\mathcal{D}^{\mathrm{same}}_{k,\mathrm{low,inc}}\Big],$$

which means it is sufficient to argue

$$\mathbb{E}\Big[|q-s|-|q-m|\;\Big|\;(q,s)\in\mathcal{D}^{\mathrm{same}}_{k,\mathrm{low,inc}}\Big]=\mathbb{E}\Big[s-m\;\Big|\;\ell\leqslant q\leqslant s<m\Big]\leqslant-\frac{\Delta}{6}.$$

Note that

$$\begin{aligned}
&\mathbb{E}\Big[s-m\;\Big|\;\ell\leqslant q\leqslant s<m\Big]\\
&\quad=-\frac{\int_{q=\ell}^{m}\int_{s=q}^{m}(m-s)f_{\mathcal{S}}(s;q)\,\mathrm{d}s\,\mathrm{d}q}{\Pr[\ell\leqslant q\leqslant s<m]}\qquad\qquad(\mathcal{Q}\text{ is uniform})
\end{aligned}$$

$$
\begin{aligned}
&\leqslant -\frac{\int_{q=\ell}^{m} \frac{1}{m-q} \left(\int_{s=q}^{m} (m-s)\,\mathrm{d}s\right) \cdot \left(\int_{s=q}^{m} f_{\mathcal{S}}(s;q)\,\mathrm{d}s\right) \mathrm{d}q}{\Pr[\ell \leqslant q \leqslant s < m]} && \text{(Lemma 1)} \\
&= -\frac{1}{2} \frac{\int_{q=\ell}^{m} (m-q) \Pr[s \in [q,m]]\,\mathrm{d}q}{\Pr[\ell \leqslant q \leqslant s < m]} \\
&\leqslant -\frac{1}{2} \frac{\frac{2(m-\ell)}{3} \cdot \int_{q=\ell}^{m} \Pr[s \in [q,m]]\,\mathrm{d}q}{\Pr[\ell \leqslant q \leqslant s < m]} && \text{(Lemma 4)} \\
&= -\frac{1}{2} \frac{\frac{2(m-\ell)}{3} \cdot \Pr[\ell \leqslant q \leqslant s < m]}{\Pr[\ell \leqslant q \leqslant s < m]} = -\frac{\Delta}{6},
\end{aligned}
$$

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

C Additional Experimental Results

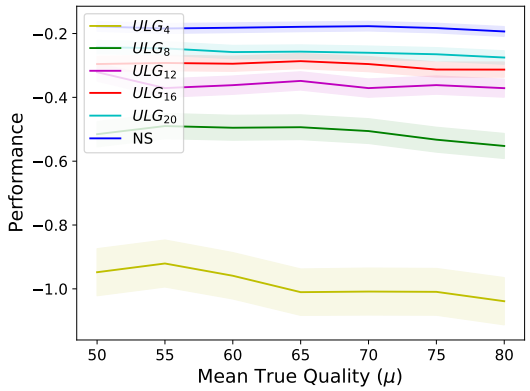
In our experiments, we compared numerical scoring to uniform letter grading schemes with $T \in \{4, 8, 12, 16, 20\}$ grades. In the main text, we presented results that show the impact of two parameters, the number of evaluations r and the motivation coefficient α_m, when one of them is varied while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

Here, we present additional experimental results, which show the impact of varying the mean μ of the true quality distribution (Figure 3), the standard deviation σ of the true quality distribution (Figure 4), and the standard deviation γ of the score distribution (Figure 5).[6]

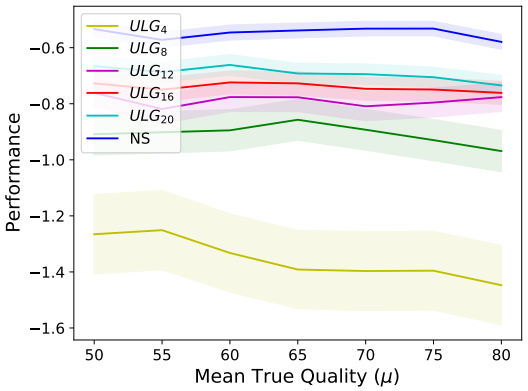
Overall, the mean true quality μ has little impact on the performance of different grading schemes. Similarly, the standard deviation σ of the true quality prior also does not significantly affect the performance of the grading schemes, but somewhat strikingly, it has a dramatic impact on the performance of ULG_4 (uniform letter grading with 4 grades).

The impact of the standard deviation γ of the score distribution is more significant, since as γ increases, the performance of the different grading schemes becomes more similar. However, we see that even for quite large values of of γ, our theoretical results seem to hold. In particularly, we see that when two evaluations are taken place, numerical scoring is better when $\alpha_m < \alpha_d$ whereas uniform letter grading is better when $\alpha_m > \alpha_d$. This observation is quite encouraging since it probably indicates that our results can be extended for cases where the score is not very well-concentrated around the true quality.

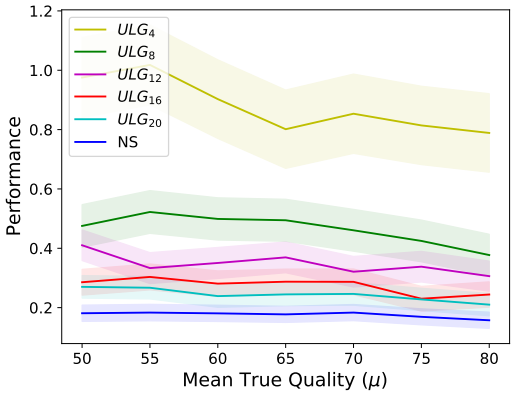
[6] Technically, these are the mean and the standard deviations of the respective underlying normal distributions before truncation.



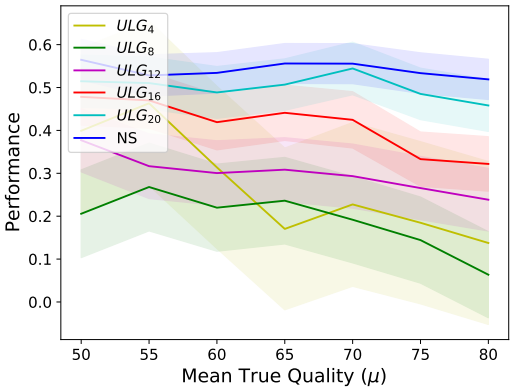
(a) $\alpha_m = 0.2, r = 2$



(b) $\alpha_m = 0.2, r = 4$

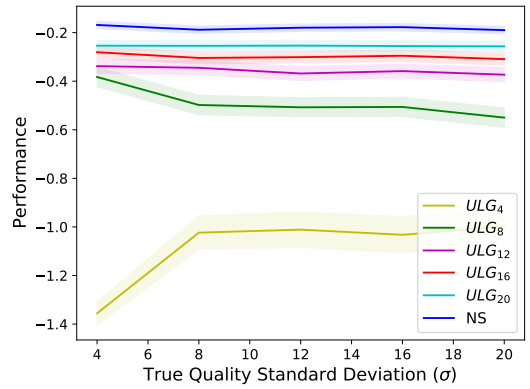


(c) $\alpha_m = 0.8, r = 2$

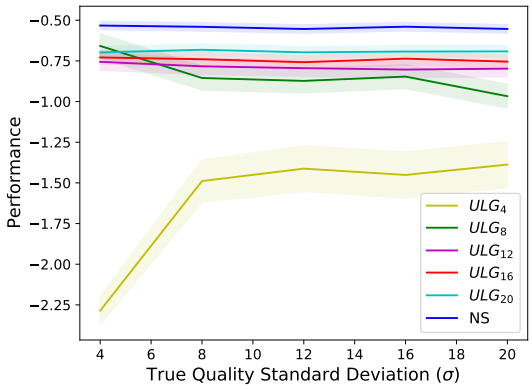


(d) $\alpha_m = 0.8, r = 4$

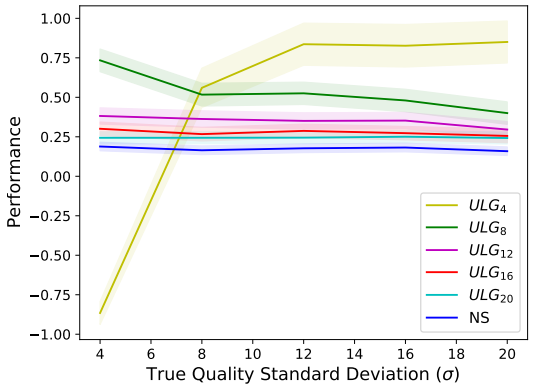
Figure 3: Performance of numerical scoring and different uniform letter grading schemes, with $\sigma = 12$, $\gamma = 1.5$ and $\alpha_d = 0.5$, over different values of μ.



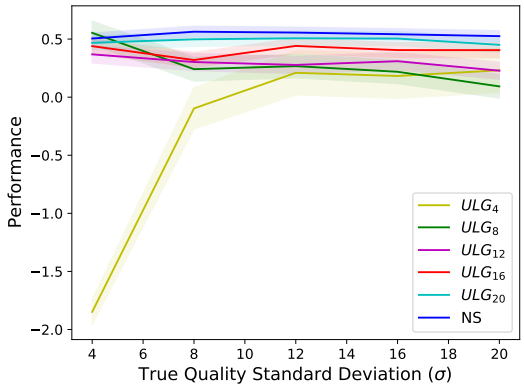
(a) $\alpha_m = 0.2, r = 2$



(b) $\alpha_m = 0.2, r = 4$

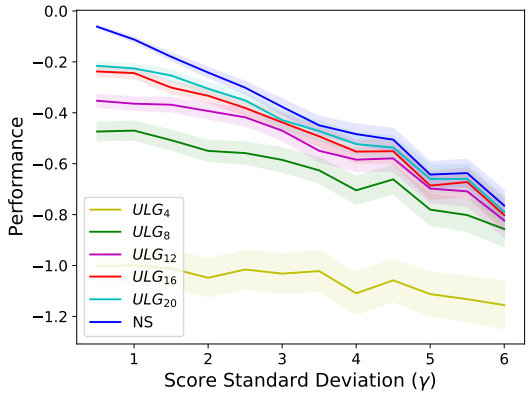


(c) $\alpha_m = 0.8, r = 2$

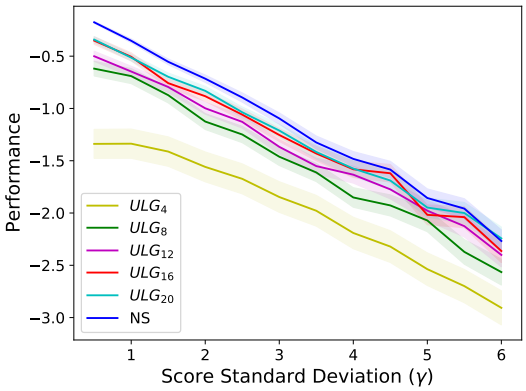


(d) $\alpha_m = 0.8, r = 4$

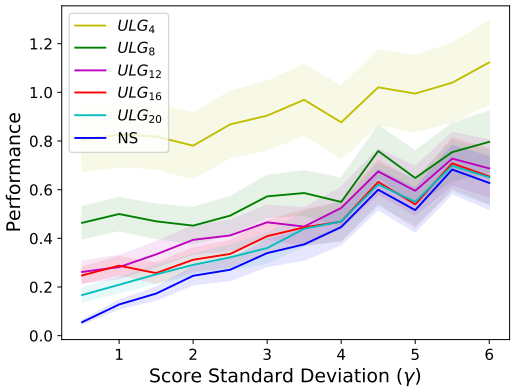
Figure 4: Performance of numerical scoring and different uniform letter grading schemes, with $\mu = 65$, $\gamma = 1.5$ and $\alpha_d = 0.5$, over different values of σ.



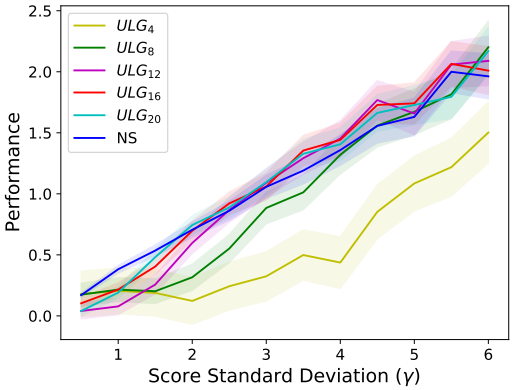
(a) $\alpha_m = 0.2, r = 2$



(b) $\alpha_m = 0.2, r = 4$



(c) $\alpha_m = 0.8, r = 2$



(d) $\alpha_m = 0.8, r = 4$

Figure 5: Performance of numerical scoring and different uniform letter grading schemes, with $\mu = 65$, $\sigma = 12$ and $\alpha_d = 0.5$, over different values of γ.