A Case for Ordinal Peer-evaluation in MOOCs

Nihar B. Shah
U. C. Berkeley
nihar@eecs.berkeley.edu

Joseph K. Bradley
U. C. Berkeley
josephkb@berkeley.edu

Abhay Parekh
U. C. Berkeley
parekh@berkeley.edu

Martin Wainwright
U. C. Berkeley
wainwrig@berkeley.edu

Kannan Ramchandran
U. C. Berkeley
kannanr@berkeley.edu

Abstract

MOOCs have been highly successful due to the ease of disseminating information: anyone with an Internet connection can watch videos of the lectures and download study material. However, they still lag far behind conventional classrooms in one critical aspect—feedback to and evaluation of the students—due to severe mismatches in the number of students enrolled and the number of experts available. One means of performing evaluation and feedback is *peer evaluation* wherein the answers submitted by a student are anonymized and provided to a set of other students to evaluate. In current peer evaluation techniques, these peer evaluators assign *cardinal* scores to the given solutions. In this paper, we explore an alternative approach to peer evaluation based on pairwise comparisons. We present evidence that such an *ordinal* approach can be significantly more robust to the lack of expertise of the evaluators, as compared to the conventional cardinal approaches. This work is a first step in understanding the trade-off between the precision of cardinal scores and the robustness of ordinal evaluations for peer grading.

1 Introduction

The advent of massive open online courses (or MOOCs) via platforms such as Coursera and EdX has enabled millions of people all over the world to gain access to quality education in a cheap and convenient manner. These courses typically have enrollments of a few thousand students, and these numbers are growing at a fast pace. The greatest advantage of MOOCs is the ease of disseminating information: anyone with an Internet connection can watch videos of the lectures and download the study material. However, MOOCs still lag far behind conventional classrooms in one critical aspect: feedback to and evaluation of the students. Due to the massive scale of these courses, it is impossible to have the instructor or the teaching assistants evaluate the thousands of answers. Moreover, since these courses are offered for free (and hopefully will continue to be), hiring paid experts for this task is not a feasible option.

The two most promising alternative evaluation techniques are auto evaluation and peer evaluation, which have been tried with varying degrees of success. Auto evaluation uses software to evaluate students' solutions to homeworks or exams. While auto evaluation techniques are well-suited for questions which are reasonably objective, such as multiple-choice questions or writing computer programs, they do not perform well for subjective questions. Consider, for instance, an essay from a literature class or a smartphone app design from a class on human-computer interaction. These assignments cannot be evaluated well by modern machine learning, necessitating the opinion of a human [1]. Indeed, many fields of study are intrinsically subjective.

An alternative means of performing evaluation and feedback is *peer evaluation*. In a system employing peer evaluation, the answers submitted by a student are anonymized and provided to a set of

Figure 1: An illustration of the feedback that a student would receive under a scheme of ordinal evaluations. The figure considers a question in which students are required to design a clock app for mobile phones.

other students to evaluate. A rubric provided by the instructor serves as a guideline for the students in their evaluation process. For example, Coursera employs peer evaluation for the human-computer interaction (HCI) course [2,3] in which the homework of each student is evaluated by 3 to 5 students, and the final evaluation of a student is computed as the median of these evaluations.

Despite its immense potential, peer evaluation has had only a limited acceptance in MOOCs. Only a handful of courses have attempted to employ peer evaluation. In these courses, the final evaluations obtained by the students under the peer evaluation process were often far from an expert evaluation. Such irregularities have lead to a severe opposition to the employment of peer evaluation in MOOCs [4]. The reasons are rather obvious. A significant fraction of students enrolled in the MOOCs are not serious about the course [5], let alone about evaluating their peers' works. The high attrition rates of students also result in complex dynamics in the system. Even when students mean well, lack of expertise and the presence of biases results in significant noise in their evaluations. The system is also faced with additional soft constraints, such as limits on the number of solutions that a student can be asked to evaluate, or existence of a very small pool of expert evaluators who need to be used smartly.

This paper considers the problem of designing peer evaluation schemes. Conventional evaluation schemes typically follow the *cardinal* approach of assigning a *score* to each student (e.g., the HCI course at Coursera). In this paper, however, we argue for the case of an alternative *ordinal* (comparative) approach towards peer evaluation. In such an approach, each student-evaluator will be given some *pairs* of solutions, and in each pair he/she must choose the 'better' solution. The pairs may comprise the solutions provided by other students, or solutions that have been previously calibrated by experts. These comparisons will then be aggregated by a system which assigns a final evaluation to each student. In what follows, we elaborate on this approach, and discuss its possible merits and demerits as compared to the traditional cardinal approach.

We believe that the cardinal and ordinal approaches both have value but differ in effectiveness depending on the problem being evaluated. We demonstrate a simple task on which people are much more capable at ordinal evaluation, yet we are also aware that simple pairwise comparisons ignore information present in cardinal scores. Our work takes a first step in understanding ordinal peer grading. In the future, a combined approach might balance the robustness of ordinal evaluation with the precision of cardinal scores.

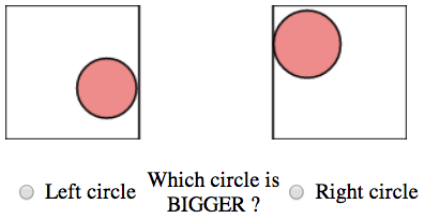
The rest of the paper is organized as follows. Section 2 compares the cardinal and ordinal approaches towards peer evaluation. Section 3 discusses our initial work at modeling ordinal peer evaluations and presents some preliminary results. Section 4 provides concluding comments.

2 Cardinal vs. Ordinal Evaluations

The traditional method of evaluation has been cardinal in nature, but such evaluations have been typically been performed by experts. It is unclear whether this approach is robust to the lack of expertise. In this section we argue for an alternative ordinal means of evaluation, and provide a qualitative comparison of the two methods.

Error in Ordinal	12.89%
Error in Cardinal	17.45%
Additional Ties in Cardinal	0.00%

(a) Estimating age of people from photographs



Error in Ordinal	6.54%
Error in Cardinal	13.50%
Additional Ties in Cardinal	15.69%

(b) Estimating areas of circles

Table 1: Screenshots and error-rates of two experiments comparing cardinal and ordinal tasks on Amazon Mechanical Turk. (The screenshots show only the ordinal part of the experiment.)

2.1 Calibration Issues

Every student has inherent biases which may vary with time or depend on the quality of the other answers evaluated. A student may be conservative and always assign moderate scores, while another student may have a tendency to inflate evaluations to the extremes. These inconsistencies are hard to learn with sample sizes as small as 5 or 10 that one would typically get in a peer evaluation setup. Furthermore, these inconsistencies vary with time, thus making them harder to learn and model. On the other hand, such inconsistencies are automatically eliminated (by design) in the ordinal setup. As illustrated below by means of practical experiments, people can be significantly more competent at performing ordinal evaluations than cardinal ones.

2.2 Experimental Results on Accuracy in Cardinal vs. Ordinal Evaluations

One could obtain ordinal evaluations from cardinal ones: after collecting cardinal evaluations, any pair of solutions can be compared using the cardinal scores. Such an argument suggests that an ordinal approach does not provide any additional data, and in fact leads to a loss of information. In this section, we present results from some experiments we performed using the Amazon Mechanical Turk crowdsourcing platform that suggest quite the opposite. The experiments reveal that when evaluations are performed by humans, ordinal evaluations contain *significantly less noise* than cardinal evaluations.

(a) Estimating age of people from photographs: Each task given to the workers required estimating the age of 10 people whose photographs were shown. In the ordinal setup, the workers had to choose the older person from pairs of pictures, while the cardinal setup required the workers to enter the estimated age. A total of 100 workers performed the tasks.

(b) Estimating areas of circles: Each task given to the workers comprised 25 questions. In the cardinal version of the tasks, for each question, the worker was shown a circle in a bounding box, and the worker was required to identify the percentage of the box's area that the circle occupied. In the ordinal version, the worker was shown two circles in separate, identical bounding boxes, and the worker was required to identify the circle that occupied a larger percentage of area in its respective box. The bounding box was 200 pixels wide and 200 pixels high, and the radius of the circle was chosen as $30 * Beta(15, 3)$, where $Beta$ denotes the Beta distribution. The cardinal answers were then converted to an ordinal form by choosing pairs of questions and looking at which circle was given a higher value. A total of 50 workers performed the tasks.

Table 1 shows the results. Converting cardinal answers to ordinal answers results in a significantly higher error rate than directly asking for ordinal evaluations.

2.3 Ease of Evaluation

The problem of peer evaluation in MOOCs requires dealing with not only the inconsistencies due to lack of expertise, but also with scarcity of resources. It is practically infeasible to ask each student

to devote too much of his/her time and effort in the peer evaluation tasks. It is thus of interest to maximize the amount of information that can be gathered under a limited effort by the students, and to design a peer evaluation process that does not fatigue the student-evaluators. To this end, it is fairly well known [6, 7] that humans often find it significantly easier to compare than score. In the setup of peer evaluation, an evaluator may find the task of providing a simple comparison to be much easier than providing a precise numerical score. An ordinal approach would then allow for the collection of more evaluations for the same level of effort, as compared to cardinal evaluations.

2.4 Concrete Constructive Feedback

Given that the evaluators are not experts, we posit that ordinal feedback, as shown in Fig. 1, may often be more desirable to the student as compared to comments on an absolute scale. Since humans are better at comparisons, the evaluator may be able to offer more insightful comments on the positive and negative aspects of the student's solution, and make these points more concrete by comparing with another solution.[1] Such a pointed evaluation will also guide the student in understanding precisely how he/she can improve his/her solution.

3 Modeling and Inference in Ordinal Peer-evaluation

In this section, we describe preliminary work on modeling and inference in ordinal peer evaluation. After discussing reasonable modeling assumptions, we give a negative result showing that any model matching these assumptions will require non-convex optimization for learning. We then propose a simple model for estimating scores from pairwise comparisons. Despite the non-convexity of the model, initial results on synthetic and real data show promise.

3.1 Modeling and Inference in Ordinal Peer evaluation

We give an axiomatic discussion of models for ordinal peer evaluation. We assume that each of n students $i \in [n]$ has an inherent skill $w_i \in \mathbb{R}$ which we wish to infer using peer comparisons. The peer comparisons let us indirectly observe the skills in two ways: the probability of j beating ℓ depends on their relative skill, and the probability of a correct comparison (where the more skilled student wins) depends on the skill of the peer evaluator.

The peer comparison mechanism works as follows. Pick a set of three students (i, j, ℓ). Anonymize the answers of students j and ℓ, and ask evaluator i which of the two answers is better. Repeat this process, choosing the students such that every answer is evaluated by some minimum number of students and such that no student performs more than some maximum number of comparisons.[2]

We can now outline generative models for comparisons. For students i, j, ℓ, we write $(i : j > \ell)$ to denote the event of evaluator i rating j above ℓ. We express a generative model using the probability $P(i : j > \ell | w_i, w_j, w_\ell)$ of observing the event $(i : j > \ell)$ conditioned on the inherent skills w_i, w_j and w_ℓ of i, j and ℓ respectively.

Note that $P(i : \ell > j | w_i, w_j, w_\ell) = 1 - P(i : \ell > j | w_i, w_\ell, w_j)$. Also note that the model may additionally depend on parameters extraneous to our current discussion (e.g., a bias or malice on the part of the grader), and this is discussed later in Corollary 2.

What properties should P have? We state three intuitive axioms:

Axiom 1 (Monotonicity with respect to grading ability) The function $P(i : j > \ell | w_i, w_j, w_\ell)$ must be non-decreasing in w_i if $w_j > w_\ell$ and non-increasing in w_i if $w_j < w_\ell$.

Axiom 2 (Monotonicity with respect to answer quality) The function $P(i : j > \ell | w_i, w_j, w_\ell)$ must be non-decreasing in w_j and non-increasing in w_ℓ.

Axiom 3 (Dependency on grader) There exists w_i, w_i', w_j, w_ℓ such that $P(i : j > \ell | w_i, w_j, w_\ell) \neq P(i : j > \ell | w_i', w_j, w_\ell)$.

[1] Note that the other solution chosen for comparison is picked cleverly by the system; the solution may be that provided by a student or may be one that has been calibrated by an expert.

[2] We assume that the set of triplets $\{(i, j, \ell)\}$ chosen for the comparisons is provided a priori to the algorithm. The problem of evaluator assignment is important, but we omit discussion due to lack of space.

Now consider the problem of inferring the skills $\mathbf{w} := [w_1, \ldots, w_n]$, given a dataset of peer comparisons, using maximum-likelihood estimation. To permit efficient optimization for inference, we would like our probabilistic model $P(i : j > \ell | w_i, w_j, w_\ell)$ to be log-concave in $\mathbf{w}$. Unfortunately, as the following result shows, there exists no log-concave function that satisfies our axioms. The proof of this result is provided in the appendix.

Theorem 1 *Suppose $P(i : j > \ell | w_i, w_j, w_\ell)$ satisfies axioms 1, 2 and 3. Then for any monotonically strictly increasing function $m : [0, 1] \to \mathbb{R}$, the function $f : \mathbb{R}^3 \to \mathbb{R}$ defined as*

$$f(w_i, w_j, w_\ell) = m(P(i : j > \ell | w_i, w_j, w_\ell))$$

cannot be concave.

Corollary 1 *$P(i : j > \ell | w_i, w_j, w_\ell)$ cannot be log-concave.*

Corollary 2 *A more generic model for the grading process may also try to incorporate additional parameters such as the bias or the maliciousness of any grader. In that case, the model considered in Theorem 1 is simply a restriction of the generic model onto a (convex) subset of its domain (e.g., in the subspace of zero bias and no malice). Consequently, the more generic model also cannot be concave.*

Corollary 3 *Theorem 1 holds even if the model restricts w_i's to belong to some interval, If one also tries to model a random grader or a perfect grader.*

This result indicates that optimization may always be difficult when modeling ordinal peer evaluation. However, our initial empirical results in the next section are encouraging. In particular, for certain models or parameter regions, optimization may be tractable. We are currently working on understanding these questions, with the goal of proving strong guarantees for optimization.

3.2 Refereed Bradley-Terry-Luce (RBTL) Model

We generalize the classical Bradley-Terry-Luce (BTL) model [8, 9] to incorporate the notion of a peer referee. The BTL model says that when two entities j and ℓ with respective skills w_j and w_ℓ are compared, j wins with probability:

$$P_{BTL}(j > l) \quad = \quad \frac{1}{1 + \exp\left(-(w_j - w_\ell)\right)} \qquad \text{(BTL model)} \tag{1}$$

We incorporate the idea of peer evaluation by scaling this probability with a function of the evaluator's skill, giving the Refereed BTL (RBTL) model:

$$P(i : j > \ell) \quad = \quad \frac{1}{1 + \exp\left(-g_i(w_j - w_\ell)\right)} \qquad \text{where } \; g_i = aw_i + b \qquad \text{(RBTL model)} \tag{2}$$

for some parameters a and b. The *evaluation ability* g_i increases linearly with the evaluator's skill w_i. A skilled evaluator with $g_i \gg 0$ will likely predict $j > \ell$ iff $w_j > w_\ell$; a malicious evaluator with $g_i < 0$ will tend to do the opposite; and a random evaluator with $g_i = 0$ will choose each of j and ℓ with equal probability. Observe that (2) reduces to the original BTL model when $a = 0, b = 1$.

Given a set of comparisons $\{(i : j > l)\}$, we can write out the data log-likelihood as a function of $\mathbf{w}$, a, b. As expected from Theorem 1, this function is not jointly convex in its parameters. We discuss empirical convergence on synthetic data in the next section.

Reasonable parameter regions: Note that the RBTL model only fits Axioms 1 and 2 if $g_i > 0$, $\forall i$. Axioms 1 and 2 thus constrain our model parameters; if we posit a generative model of the skills w_i, then we can write down constraints for parameters a, b. For example, if $w_i \sim \mathcal{N}(\mu, \sigma^2)$ i.i.d., and $\Phi(\cdot)$ is the standard normal CDF, then we get for each i:

$$P(\text{malicious}) = \Phi(-\tfrac{\mu + b/a}{\sigma}) \, . \tag{3}$$

One could imagine many models of ordinal peer evaluation; we make no claim that the RBTL model is ideal. Yet exploring the RBTL model helps us to understand two key questions:

- How hard is optimization for a model with reasonable parameters encoding our axioms?
- How strong is the relation between student skill w_i and evaluator ability g_i?

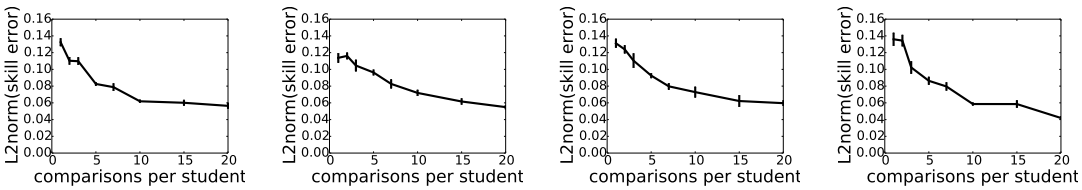


Figure 2: **Skill Estimation Error vs. Number of Comparisons per Student.** Maximum-likelihood estimation for the gBLT model with over 50 students, $b = 1$ and various values of a. A larger value of a implies a stronger dependence between the skill w_i and grading ability g_i of each student; $a = 0$ gives the original BTL model. Error bars show stderr from 5 trials.

3.3 Simulations

To understand the parameterization of the RBTL model, and to test the difficulty of optimization, we ran tests on data simulated from the RBTL model in (2) for various settings of parameters a, b. We generated skills $w_i \sim \mathcal{N}(0, 1)$ and comparisons using the RBTL model. We then trained the RBTL model to estimate the skills $\mathbf{w}$, fixing a, b to the ground truth. We used stochastic gradient to optimize the data log likelihood, with L2-regularization set according to the prior from which we generated $\mathbf{w}$. We set the stochastic gradient step size to $1/\sqrt{T}$ on iteration T.

In short, optimization did not pose a significant problem. To give an idea of sample complexity, we have plotted results in Figure 2 for selected models. The error in skill estimates drops quickly as each grader makes more comparisons, but note that many parameter settings still require 10 to 20 comparisons per grader to achieve low error. Interestingly, even though we expect about 8 malicious graders when $a = 1$, estimation is still easy. We posit that large a makes outcomes more certain (i.e., more signal per comparison), even while increasing the number of malicious graders.

3.4 Experiments on Real Data

We tested on data from the third offering of Human Computer Interaction (HCI) on Coursera, taught by Prof. Scott Klemmer (then at Stanford). See [3] for details on the peer evaluation system. In our dataset of peer evaluations for the first English homework assignment, we had 1879 students and 7242 numerical peer evaluations.

We compared our RBTL model with the original BTL model and with median prediction (using the median of a student's grades as the estimated grade). We tested using 4-fold cross validation on the peer grades, training on 3/4 of the peer grades and testing on the held-out 1/4 of the grades. We created comparison data from peer evaluations by taking each evaluator i and comparing the scores i gave to each pair of peers (j, ℓ). This produced 16,310 peer comparisons.

We fit the RBTL model by fixing $b = 1$ and estimating $\mathbf{w}$ and a by maximizing training data log likelihood, with regularization. We used alternating block coordinate descent, alternating between (a) fixing a and optimizing $\mathbf{w}$ with stochastic gradient and (b) fixing $\mathbf{w}$ and optimizing a via a line search.

We regularized $\mathbf{w}$ with the equivalent of a $\mathcal{N}(1, \sigma^2)$ prior, and we applied L2 regularization to a with parameter λ_a. We chose σ^2 and λ_a via 3-fold cross-validation. We fit the BTL model analogously, except that a was fixed at 0. Cross validation chosen $(\sigma^2, \lambda_a) = (0.5, 100)$ for the RBTL and $\sigma^2 = 100$ for the BTL model.

Training gave an average of $a = 3.38$ for the RBTL model. Recall our discussion of malicious evaluators in (3). Under our prior for $\mathbf{w}$, these settings for σ^2 and (a, b) imply few malicious evaluators: $P(\text{malicious}) \approx 0.0334$.

We evaluated the error by comparing each of our three model's predictions of pairwise comparisons within the held-out test data. For a given peer comparison $(i : j > \ell)$, our error metric `0-1 Error` has value 0 if the model assigns a higher score to student j than ℓ, and value 1 otherwise. The error metric `ProbError` has value $|P_{model}(i : j > \ell) - 1|$ for every comparison $(i : j > \ell)$.

	RBTL	BTL	median
mean `0-1 Error`	**0.241671**	0.272198	**0.241407**
stderr `0-1 Error`	0.00538454	0.0107655	0.00509211
mean `ProbError`	**0.285829**	0.44362	–
stderr `ProbError`	0.0058381	0.00130699	–

Figure 3: **Results on HCI Peer evaluations.** The RBTL, BTL, and median model errors for predicting the results of pairwise comparisons of students in held-out peer grades. Discarding the numerical data in peer grades hurts the performance of the BTL model, which does worse than the median prediction. Including grading ability allows the RBTL model to make up for this loss in performance.

We show our initial results in Figure 3. The BTL model does the worst; the RBTL model and the median prediction have similar performance. Modeling grading ability captures an important effect which significantly improves the performance of ordinal models, in terms of both predicting outcomes of pairwise comparisons (`0-1 Error`) and predicting the probability of outcomes (`ProbError`). The large magnitude of $a = 3.38$ selected via cross-validation, relative to $b = 1$, also highlights the value of modeling grading ability.

It is encouraging that the RBTL model can match the performance of a numerical method, even though converting numerical scores to comparisons discards information. The arguments of Section 2 hint at the possibility of obtaining higher-quality ordinal evaluations if the evaluators are directly asked to perform ordinal feedback, as opposed to converting cardinal scores into ordinal evaluations. We posit that, if peer grades were collected as comparisons, our RBTL model would have improved performance.

4 Conclusions

We posit an ordinal approach to peer-evaluation in MOOCs, which we argue is robust to the lack of expertise among the graders. We also present initial work on modeling and analyzing such an ordinal setting, and obtain encouraging results from preliminary experiments on real and synthetic data.

Acknowledgements

The authors would like to thank Jonathan Huang for allowing us to run experiments on the peer-evaluation data from the HCI courses at Coursera.

References

[1] Professionals against machine scoring of student essays in high-stakes assessment. `http://humanreaders.org/petition/index.php`.

[2] Chris Piech, Jonathan Huang, Zhenghao Chen, Chuong Do, Andrew Ng, and Daphne Koller. Tuned models of peer assessment in MOOCs. In *International Conference on Educational Data Mining*, 2013.

[3] C. Kulkarni, K. Pang-Wei, H. Le, D. Chia, K. Papadopoulos, D. Koller, , and S. R. Klemmer. Scaling self and peer assessment to the global design classroom. In *CHI*, 2013.

[4] Jonathan Rees. Peer grading cant work. `http://www.insidehighered.com/views/2013/03/05/essays-flaws-peer-grading-moocs`.

[5] Katy Jordan. MOOC completion rates. `http://www.katyjordan.com/MOOCproject.html`.

[6] William Barnett. The modern theory of consumer behavior: Ordinal or cardinal? *The Quarterly Journal of Austrian Economics*, 6(1):41–65, 2003.

[7] Neil Stewart, Gordon DA Brown, and Nick Chater. Absolute identification by relative judgment. *Psychological review*, 112(4):881, 2005.

[8] Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, pages 324–345, 1952.

[9] R Duncan Luce. Individual choice behavior, a theoretical analysis. *Bull. Amer. Math. Soc. 66 (1960), 259-260 DOI: http://dx. doi. org/10.1090/S0002-9904-1960-10452-1 PII*, pages 0002–9904, 1960.

Appendix

4.1 Proof of Theorem 1

Let us first assume that the function $P(i : j > \ell | w_i, w_j, w_\ell)$ depends only on the values w_i and $(w_j - w_\ell)$. Let $g = w_i$ and $z = w_j - w_\ell$. Define function $s : \mathbb{R}^2 \to \mathbb{R}$ as

$$s(g, z) := m(P(i : j > \ell | w_i, w_j, w_\ell))$$

where $m : [0, 1] \to \mathbb{R}$ is an arbitrary monotonically strictly increasing function. We shall first show that the function s cannot be concave. This is proved via a contradiction argument, for which we assume that there is indeed such a function s which is concave.

Firstly, since $P(i : j > \ell | w_i, w_j, w_\ell) + P(i : \ell > j | w_i, w_\ell, w_j) = 1$, it must be that for any $g_1,\ g_2,\ z_1 \geq 0,\ z_2 \geq 0$,

$$s(g_1, z_1) > s(g_2, z_2) \iff s(g_1, -z_1) < s(g_2, -z_2) \tag{4}$$

$$s(g_1, z_1) < s(g_2, z_2) \iff s(g_1, -z_1) > s(g_2, -z_2) \tag{5}$$

$$s(g_1, z_1) = s(g_2, z_2) \iff s(g_1, -z_1) = s(g_2, -z_2)\ . \tag{6}$$

In addition, if $z = 0$ (i.e., if $w_j = w_\ell$) then $P(i : j > \ell | w_i, w_j, w_\ell) = \frac{1}{2}\ \forall w_i$. Defining a constant h as

$$h = m\left(\frac{1}{2}\right)\ ,$$

we have that for any $g \in \mathbb{R}$,

$$s(g, 0) = h\ . \tag{7}$$

Further, from the Axioms 1 and 2, we have that for any $g_1,\ g_2 \geq g_1,\ z_1 \geq 0,\ z_2 \geq z_1$,

$$s(g_1, z_1) \leq s(g_2, z_2) \tag{8}$$

$$s(g_1, -z_1) \geq s(g_2, -z_2)\ . \tag{9}$$

The proof proceeds in five steps:

- Step 1: For any g_0, $z_1 \geq 0$, $z_2 > z_1$, it must be that $s(g_0, z_1) < s(g_0, z_2)$.
- Step 2: For any g_0, $z_0 > 0$, it must be that $s(g_0, z_0) > h$.
- Step 3: For any g_1, $g_2 < g_1$, $z_1 \geq 0$, $z_2 > z_1$, it must be that $s(g_1, z_1) < s(g_2, z_2)$.
- Step 4: The size of the set $\mathcal{Z} := \{z \in \mathbb{R} | \exists g_1, g_2 \text{ such that } s(g_1, z) \neq s(g_2, z)\}$ is countable.
- Step 5: The set $\mathcal{Z}$ is empty.

Step 5 thus makes the entire generative model completely independent of the grader quality, thus leading to a violation of Axiom 3.

Proof of Step 1: From (8) we know that $s(g_0, z_1) \leq s(g_0, z_2)$. Now suppose $s(g_0, z_1) = s(g_0, z_2)$. Consider any $z_3 > z_2$. Then

$$s(g_0, -z_3) \leq s(g_0, -z_2) = s(g_0, -z_1) \leq s(g_0, 0) \leq s(g_0, z_1) = s(g_0, z_2) \leq s(g_0, z_3)\ .$$

A strict inequality $s(g_0, 0) < s(g_0, z_1)$ would, however, result in

$$s(g_0, -z_2) = s(g_0, -z_1) < s(g_0, 0) < s(g_0, z_1) = s(g_0, z_2)$$

which contradicts the concavity of function s. A strict inequality $s(g, z_2) < s(g, z_3)$ will result in

$$s(g_0, -z_3) < s(g_0, -z_2) = s(g_0, -z_1) \leq s(g_0, 0) \leq s(g_0, z_1) = s(g_0, z_2) < s(g_0, z_3)\ .$$

which also contradicts the concavity of function s. As a result, we must have $\forall z \in \mathbb{R}$,

$$s(g_0, z) = s(g_0, 0) \tag{10}$$

$$= h\ . \tag{11}$$

Now due to Axiom 3, there must exist some $g_4 \in \mathbb{R}$ and $z_4 > 0$ such that $s(g_4, z_4) \neq h$. From (9), it must also be that $s(g_4, -z_4) \leq h$. It follows that

$$s(g_4, -z_4) < h \tag{12}$$

$$= \frac{1}{2}s(g_0, -2z_4) + \frac{1}{2}s(2g_4 - g_0, 0) \tag{13}$$

$$\leq s(g_4, -z_4) \tag{14}$$

where (13) results from $s(g_0, -z_4) = h$ and $s(2g_4 - g_0, 0) = h$ and (14) is a consequence of the concavity of s. This leads to a contradiction, thereby proving Step 1.

Proof of Step 2: From Step 1 we have

$$\begin{aligned} s(g_0, z_0) &> s(g_0, 0) && (15)\\ &= h\ . && (16)\end{aligned}$$

Proof of Step 3: Suppose $s(g_1, z_1) \geq s(g_2, z_2)$. From Step 2, we have

$$\begin{aligned} s(g_1, z_1) &> h && (17)\\ &= s\left(g_1 \frac{z_2}{z_2 - z_1} - g_2 \frac{z_1}{z_2 - z_1}, 0\right) && (18)\end{aligned}$$

It follows that

$$s(g_1, -z_1) \leq s(g_2, -z_2) \tag{19}$$

and

$$s(g_1, -z_1) < s\left(g_1 \frac{z_2}{z_2 - z_1} - g_2 \frac{z_1}{z_2 - z_1}, 0\right) \tag{20}$$

and hence

$$\begin{aligned} s(g_1, -z_1) &< \frac{z_1}{z_2} s(g_2, -z_2) + \left(1 - \frac{z_1}{z_2}\right) s\left(g_1 \frac{z_2}{z_2 - z_1} - g_2 \frac{z_1}{z_2 - z_1}, 0\right) && (21)\\ &\leq s(g_1, -z_1) && (22)\end{aligned}$$

where the final step is due to the assumed concavity of s. This is a contradiction. This proves Step 3.

Proof of Step 4: Consider any $g_0 \in \mathbb{Z}$. Suppose there exists some $z_0 > 0$ such that $s(g_0 + 1, z_0) \neq s(g_0, z_0)$. From (8), it follows that

$$s(g_0 + 1, z_0) > s(g_0, z_0)\ . \tag{23}$$

Take any $\epsilon > 0$. From Step 3, it also follows that

$$s(g_0 + 1, z_0) < s(g_0, z_0 + \epsilon)\ . \tag{24}$$

This must be true for any arbitrarily small value of ϵ. Define a function $s_0 : \mathbb{R} \to \mathbb{R}$ as $s_0(z) = s(g_0, z)$. It follows that the function $s_0(z)$ must be discontinuous at $z = z_0$. Furthermore, we know from (8) that $s_0(z)$ is non-decreasing in z. As a result, the function s_0 can have only a countable number of discontinuities. It follows that the number of points z_0 where $s(g_0 + 1, z_0) > s(g_0, z_0)$ must be countable. An identical argument applies to the case when $z_0 < 0$. Thus the size of the set $\mathcal{Z}_{g_0} := \{z \in \mathbb{R} | s(g_0 + 1, z) > s(g_0, z)\}$ is countable. Now,

$$\mathcal{Z} = \bigcup_{g \in \mathbb{Z}} \mathcal{Z}_g\ . \tag{25}$$

Since $\mathcal{Z}$ is a union of a countable number of countable sets, it itself is countable. This proves Step 4.

Proof of Step 5: Consider any $z_1 > 0$ such that $z_1 \in \mathcal{Z}$. Since the size of $\mathcal{Z}$ is countable, there must exist some $z_0 < z_1$ and $z_2 > z_1$ such that $z_0 \notin \mathcal{Z}$ and $z_2 \notin \mathcal{Z}$. Hence, for all $g \in \mathbb{R}$, $s(g, z_0) = s(0, z_0)$ and $s(g, z_2) = s(0, z_2)$. From (8), it also follows that

$$s(g, z_0) \leq s(g, z_1) \leq s(g, z_2)$$

and hence

$$s(0, z_0) \leq s(g, z_1) \leq s(0, z_2)$$

for all $g \in \mathbb{R}$. Define function $s_1 : \mathbb{R} \to \mathbb{R}$ as $s_1(g) = s(g, z_1)$. It follows that the function s_1 is bounded from above as well as bounded from below, and furthermore is monotonic (from (8)) and concave (due to the assumed concavity of function s). This mandates s_1 to be a constant-valued function. This contradicts the claim of having $z_1 \in \mathcal{Z}$. An identical argument holds for the case of $z_1 < 0$. This proves Step 5.

As mentioned previously, a consequence of Step 5 is that $s(g_1, z) = s(g_2, z)$ for all $g_1 \in \mathbb{R}$, $g_2 \in \mathbb{R}$ and $z \in \mathbb{R}$. This means that the generative model is independent of the grader quality, thus violating Axiom 3, thereby causing a contradiction.

Finally, we return to the general function $f(w_i, w_j, w_\ell)$. If this function is concave, so is the equivalent function where w_j and w_ℓ are replaced by an invertible linear transformation $(w_j - w_\ell)$ and $(w_j + w_\ell)$. However, for any fixed value of $(w_j + w_\ell)$, this is simply the function $s(g, d)$ which we know cannot be concave. ■

4.2 Proof of Corollary 3

Suppose w_i is bounded from below (say, by L). Then, in order to model a grader who grades randomly, one must have $P(i : j > \ell | w_i = L, w_j, w_\ell) = \frac{1}{2}$. It follows that for any $z_1 > 0$ and $z_2 > z_1$,

$$s(L, z_1) = s(L, z_2) \ . \tag{26}$$

This contradicts Step 1 in the proof of Theorem 1.

Suppose w_i is bounded from above (say, by U). Then, in order to model a grader who grades perfectly, one must have $P(i : j > \ell | w_i = U, w_j, w_\ell) = 1$ whenever $w_j > w_\ell$ and 0 whenever $w_j < w_\ell$. It follows that for any $z_1 > 0$ and $z_2 > z_1$,

$$s(U, z_1) = s(U, z_2) \ . \tag{27}$$

This contradicts Step 1 in the proof of Theorem 1.

Note that if w_i is bounded from below but not from above, then the proof of Theorem 1 itself goes through.