Two is better than one: optimizing binary rating scales for online platforms

Liren Shan*
Toyota Technological Institute at Chicago

Alec Sun*
University of Chicago

Jingyan Wang*
Toyota Technological Institute at Chicago

Abstract

Online platforms often struggle with inflated, uninformative user ratings. This issue can be mitigated by better design of the data collection mechanism, for example through changing the phrasing of user questions. Previous work optimizes a single rating scale to best distinguish the inherent qualities of items under binary feedback. In this paper, we demonstrate the power of utilizing different rating scales for different users. First, we show that a designer can always split a single rating scale into two to improve item rankings. Next, we formally characterize the optimal design of multiple scales as a series of 0-1 step functions whose threshold values are evenly spaced. Finally, we consider an adaptive model where subsequent rating scales can depend on the responses from previous users and items are ranked by their full sequence of responses, and show that adaptivity leads to exponentially greater distinguishing power. All in all, our theoretical results show that using multiple rating scales significantly improves item rankings on platforms.

1 Introduction

User data is the backbone of online platforms. Ratings are collected, for example, for videos on YouTube, products on Amazon, businesses on Yelp, freelancers on Taskrabbit, and so on. Such rating data provides valuable information to users to inform their decision making [Rachmiani et al., 2024]. Moreover, recommendation and personalization algorithms rely on user data to estimate the underlying quality of products, learn user preferences, and make future recommendations. However, ratings are often uninformative in practice, largely due to *rating inflation*, a tendency for users to give very high ratings to most items [Aziz et al., 2023]. For example, 74% of guest reviews for Airbnb properties receive a perfect rating of 5 stars [Fradkin et al., 2021]. The percentage of positive ratings for an eBay seller has a median of 100% and a mean of 99.3% [Nosko and Tadelis, 2015]. Under inflation, it becomes difficult to distinguish the quality of items. Users need to rely on small differences that are often noisy in nature, for example, comparing restaurants with scores of 4.8 and 4.9. Beyond inflation, another common pattern for user ratings is a bimodal shape, where ratings tend to be either very high or very low to convey that users were either very satisfied or very dissatisfied. These ratings again become noisy and uninformative.

Prior work such as Garg and Johari [2019] has made the observation that the distribution of user ratings is not fixed, but instead depends on the design of data elicitation. The language used in asking a question influences the user responses. For example, for binary questions, let us compare the two questions "did you have a positive experience?" and "was your experience the best you have ever had?". It is natural to anticipate, and empirically validated [Garg and Johari, 2019], that the former

*Authors are listed in alphabetical order.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

question induces users to apply a more lenient calibration and provide more positive ratings. By designing these questions carefully, the platform mitigates rating inflation, increases data quality, makes more effective future recommendations, and thus improves the user experience on the platform.

We follow Garg and Johari [2019] and consider binary responses because they are simple to implement and consistent with the well-known fact that people are unable to use fine-grained numerical values to indicate subtle differences accurately [Miller, 1956]. The design of binary rating scales is mathematically formalized as follows. We represent a rating scale by a nondecreasing function $\beta : [0, 1] \to [0, 1]$, mapping the item's inherent quality to the likelihood that a user provides a positive rating. Specifically, an item with true quality $\theta \in [0, 1]$ receives a positive rating from a user with probability $\beta(\theta)$. Going back to the previous example questions, "did you have a positive experience?" may be represented by $\beta(\theta) \approx \theta$, where the probability of an item being rated positively smoothly increases in its quality θ. In contrast, the question "was your experience the best you have ever had?" corresponds to a function β whose value is close to 0 unless θ is close to 1.

Under this model, the goal is to compute the optimal binary scale β under a certain objective, to be introduced formally in Section 3, for distinguishing the quality of items. For a single binary rating scale, Garg and Johari [2019] construct an optimization problem that can be solved numerically to compute the scale that maximizes the speed at which items of different qualities can be distinguished from each other. This optimal scale is then used to elicit binary ratings from all users on the platform.

Departing from prior work, in this paper we are motivated by the fact that it is feasible and convenient to apply different rating scales to different users, for example by asking half of the users if their experience was above average and the other half if their experience was the best they have ever had. Allowing multiple rating scales generalizes a single rating scale because one can always set all the scales to be the same. Hence using multiple scales weakly dominates using a single scale in distinguishing items, but the precise gains are a priori unclear. Our research question is therefore:

> *How should a platform design multiple scales, and what is the quantitative gain compared to a single scale?*

Overview of results. We characterize the gain from using multiple rating scales for data collection through the following theoretical results:

- Given a single rating scale applied to all users, one can always construct a pair of rating scales, where each rating scale is applied to half of the users, such that the ranking objective under the two rating scales is at least that under the single scale, and strictly higher unless the single scale is already a 0-1 step function or a constant.
- If we allow the number of rating scales to equal the number of users, so that each user rates on a different scale, then we derive the optimal solution exactly: each user rates under a distinct 0-1 step function $f_\tau(\theta) = \begin{cases} 0 & \text{if } \theta < \tau \\ 1 & \text{if } \theta \geq \tau \end{cases}$ with threshold at τ.
- Under an *adaptive* model where users arrive sequentially, we consider scales chosen adaptively based on the rating scales and responses of all previous users, and we let the platform rank items by their full sequence of responses rather than by their number of positive ratings. We again derive the optimal solution, where each rating scale is still a 0-1 step function but with different threshold values compared to the non-adaptive setting: the scales perform a binary search for the item's quality. We show that adaptivity provides an exponential gain in the objective compared to the non-adaptive setting, and that this gain requires changing the aggregation rule in addition to the elicitation rule.

2 Related work

Rating inflation and rating system design. Ratings in online marketplaces often become inflated and uninformative, due to the pressure of not hurting sellers [Filippas et al., 2022], reciprocity between sellers and buyers [Dellarocas and Wood, 2008], and so on. The design of the rating system, through changing the verbal description of each point on the scale [Klockars and Yamagishi, 1988, Garg and Johari, 2021], has been proposed as an effective way to mitigate inflation. More quantitatively, Garg and Johari [2019] compute the binary rating scale that optimizes certain measures of rating informativeness.

Miscalibration. Ratings from people are known to be miscalibrated [Griffin and Brenner, 2004], and various computation methods are proposed to correct such miscalibration in a post-hoc fashion [Paul, 1981, Roos et al., 2011, MacKay et al., 2017]. Instead of correcting the scores collected under the inherent miscalibration of people, the design of rating systems can be viewed as changing this calibration as part of the data collection process.

Aggregating heterogeneous data in crowdsourcing. Our subsequent results demonstrate the gain from using heterogeneous scales. At a high level, the gain comes from each scale specializing in part of the information collection task. For example, one scale is effective in distinguishing items of low quality, and another scale is effective in distinguishing top items. This specialization is similar in spirit to aggregating heterogeneous human data in crowdsourcing [Dawid and Skene, 1979, Welinder et al., 2010]. Again, instead of learning the heterogeneous expertise of people from data, the rating scale allows us to endogenously induce the heterogeneity under which information collection is the most effective.

Item response theory. Item response theory [Embretson and Reise, 2013] in psychometrics concerns estimating the latent quality of students based on their responses to test questions, which is similar to our problem of ranking items based on user ratings. Our adaptive setting is also similar to adaptive testing, where test questions are selected adaptively based on a student's previous responses. Instead of choosing from a fixed bank of questions or rating scales, we characterize the optimal "response functions" to guide the design of questions.

3 Problem formulation

The platform wishes to rank a set of items where each item has a unique and unknown quality $\theta \in [0, 1]$. We assume without loss of generality that these qualities are uniformly distributed in $[0, 1]$, noting that we can reduce any distribution to this case by rescaling, so that θ is the item's quality *quantile* rather than its absolute quality. We assume that each item will accumulate a total of k ratings from users on the platform.

Rating scales. Each user's rating is binary and collected under a *rating scale*, which is a nondecreasing function $\beta(\theta) : [0, 1] \to [0, 1]$ such that an item with quality θ gets a positive rating with probability $\beta(\theta)$ and a negative rating with probability $1 - \beta(\theta)$. Ratings of different items and of different users are independent given the item qualities. Prior work [Garg and Johari, 2019] has primarily been concerned with optimizing a *single* rating scale β that distinguishes items of different qualities. For a single scale, ratings for an item of quality θ are always distributed as $\mathrm{Bern}(\beta(\theta))$ and hence the number of positive ratings after k users is $\mathrm{Binom}(k, \beta(\theta))$. Moving beyond a single scale, we consider *multiple scales*, where the rating scales $\beta(\theta)$ need not be identical across the k users. For each $i \in [k]$, we denote the rating scale used for the i-th user by β_i. The total number of positive ratings that an item with quality θ accumulates is a random variable $n_k(\theta) := \sum_{i=1}^{k} \mathrm{Bern}(\beta_i(\theta))$.

Platform objective. The platform aims to design a mechanism for collecting user ratings such that the ranking induced by the mean ratings of these items is as accurate as possible. That is, given two items with qualities $\theta_1 < \theta_2$, the platform ranks them accurately if θ_2 accumulates more positive ratings than θ_1, namely $n_k(\theta_1) < n_k(\theta_2)$. We consider an accuracy metric whose value is 1 if the two items are ranked correctly, -1 if they are ranked incorrectly, and 0 if they are tied in terms of their mean ratings. This ranking accuracy between items with qualities $\theta_1 < \theta_2$ is thus captured by the expression:

$$P_k(\theta_1, \theta_2 \mid \beta) := \mathbb{P}[n_k(\theta_1) < n_k(\theta_2)] - \mathbb{P}[n_k(\theta_1) > n_k(\theta_2)] \,.$$

The platform generally wishes to maximize $P_k(\theta_1, \theta_2 \mid \beta)$ across all pairs of item qualities $\theta_1 < \theta_2$. However, it may care about ranking certain pairs of qualities more than others, which is captured by an integrable *weight function* $w(\theta_1, \theta_2) : [0, 1] \times [0, 1] \to \mathbb{R}_{\geq 0}$ that indicates how much the platform values ranking items with qualities $\theta_1 < \theta_2$ correctly.

Example 1. *The most commonly used weight functions are the following:*

- *Kendall's τ function weights each pair of qualities uniformly, namely $w(\theta_1, \theta_2) = 1$ for all $\theta_1, \theta_2 \in [0, 1]$.*

- *Spearman's ρ function weights pairs of qualities proportional to the difference in qualities, namely $w(\theta_1, \theta_2) = \theta_2 - \theta_1$ for all $\theta_1, \theta_2 \in [0, 1]$. In words, it is more important to distinguish items with significantly different qualities than those with similar qualities.*

More generally, varying the weight function $w(\theta_1, \theta_2)$ allows the platform to encode a rich set of objectives [Tarsitano, 2009, Pinto Da Costa and A.C. Roque, 2006].

The platform seeks to maximize the following overall objective, which is obtained by integrating $P_k(\theta_1, \theta_2 \mid \beta)$ across all $\theta_1 < \theta_2$, weighting each pair of qualities (θ_1, θ_2) by $w(\theta_1, \theta_2)$, and finally normalizing by a factor of 2 since $\mathbb{P}[\theta_1 < \theta_2] = \frac{1}{2}$:

$$W_k := 2 \int_{\theta_1 < \theta_2} w(\theta_1, \theta_2) \cdot P_k(\theta_1, \theta_2 \mid \beta)\, d\theta_1 d\theta_2.$$

Note that the theoretical maximum of W_k under Kendall's τ function is 1, which occurs when all items can be perfectly distinguished from each other, namely $P_k(\theta_1, \theta_2 \mid \beta) = 1$ for all $\theta_1 < \theta_2$.

Comparing single and multiple scales. Recall that the goal of this paper is to understand the power of using different rating scales $\{\beta_i\}_{i=1}^k$ for different users rather than using a single scale β for all k ratings. To do this, we would like to compute the maximum platform objective allowing multiple rating scales, which we denote by W_k^{multiple}, and show that it is greater than the maximum objective that uses a single rating scale, which we denote by W_k^{single}. For this comparison to be meaningful, we must use the same weight function $w(\theta_1, \theta_2)$, so we use Kendall's τ function for ease of exposition. Our theoretical results also show how to generalize to other weight functions $w(\theta_1, \theta_2)$. In Section 5 we prove that $W_k^{\text{multiple}} = 1 - \frac{1}{k+1}$ for Kendall's τ function if each rating can be on its own scale, which means that using multiple scales approaches perfect accuracy at a rate of order $\frac{1}{k}$. On the other hand, we are unable to compute W_k^{single} exactly. Instead, we employ two methods to *estimate* W_k^{single}. The first method is deriving a theoretical upper bound $\bar{W}_k^{\text{single}}$ for W_k^{single}:

Proposition 2 (Upper bound on platform objective for single scale)**.** *Let $w(\theta_1, \theta_2)$ be Kendall's τ function. The best platform objective W_k^{single} for k ratings using a single scale is at most*

$$\bar{W}_k^{\text{single}} := 1 - 2^{k-1} \int_0^1 \frac{y^k}{(1+y)^{k-1}}\, dy. \tag{1}$$

The proof of Proposition 2 is in Appendix A.2. The second method is solving for the best *discretized* single scale. To do this, we split $[0, 1]$ into n equally spaced intervals and solve for the optimal step function that is constant within each interval and maximizes W_k. The results of this computation are shown in Figure 5.

Adaptive rating scales. The model above, following Garg and Johari [2019], has two components. The first is an *elicitation* rule: the k scales $\beta_1, \ldots, \beta_k$ are set by the designer in advance, so the scale shown to a user does not depend on the ratings of previous users. The second is an *aggregation* rule: an item's score is its number of positive ratings $n_k(\theta)$, and the item with the higher count is ranked higher. In Section 6 we also consider an *adaptive* model that modifies both components. For elicitation, the designer uses a scale β_1 for the first user, then uses the first user's binary rating to choose the second scale β_2, and in general chooses the scale β_j for the j-th user as a function of the ratings of the first $j-1$ users. For aggregation, the platform ranks items by their full sequence of k binary responses rather than by the number of positive responses. This sequence identifies the path that the item took through the tree of scales. We show in Section 6 that the second change is necessary for the first to be useful, and that the optimal configuration of adaptive rating scales distinguishes items at an exponentially better rate of $W_k^{\text{adaptive}} = 1 - \frac{1}{2^k}$ compared to the non-adaptive rate of $W_k^{\text{multiple}} = 1 - \frac{1}{k+1}$.

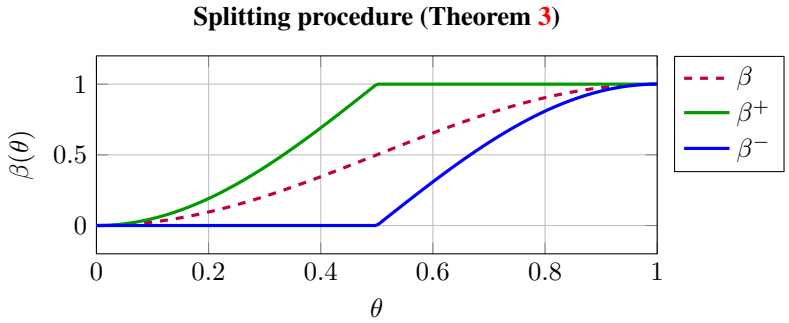


Figure 1: Splitting a single scale $\beta(\theta)$ into two scales $\beta^+(\theta) = \min\{2\beta(\theta), 1\}$ and $\beta^-(\theta) = \max\{2\beta(\theta) - 1, 0\}$ increases the platform objective W_k.

4 Local result: two is better than one

We first show a *local* improvement result: Starting with any single scale β used for an even number of ratings, we construct two scales β^+ and β^-, such that using each scale for half of the ratings improves the platform objective compared to using β for all ratings.

Theorem 3. *Let $k = 2m$ be any even integer, and let w be any nonnegative weight function. Given any single rating scale $\beta(\theta)$ for k ratings, there exist two scales*

$$\beta^+(\theta) := \min\{2\beta(\theta), 1\}, \qquad \beta^-(\theta) := \max\{2\beta(\theta) - 1, 0\}$$

such that the platform objective W_k achieved by using the scale β^+ for m ratings and the scale β^- for the other m ratings is at least that achieved by using scale β for all k ratings. Furthermore, if $w > 0$ almost everywhere, the increase in W_k is strict unless β is a 0-1 step function or a constant.

The scales β^+ and β^- are constructed by "splitting" the original scale β, visualized in Figure 1. Intuitively, a steep curve in the scale suggests that users respond very differently to a pair of items, and hence the rating scale achieves a higher accuracy. In this construction, the scale β^+ focuses on distinguishing items of low quality, by amplifying the original scale β for these items. Likewise, the scale β^- focuses on distinguishing items of high quality. This splitting preserves the expected number of positive ratings while reducing the noise in the ratings. For example, consider θ with $p := \beta(\theta) < 1/2$. The two split ratings are distributed as $\mathrm{Bern}(2p)$ and 0, whose total variance is $2p(1-2p)$, smaller than the variance $2p(1-p)$ of two independent $\mathrm{Bern}(p)$ ratings.

We also note that the splitting procedure can be performed recursively to further increase the platform objective. Starting with any number of scales, we split any one of these scales β used for $k = 2m$ ratings into two scales each used for m ratings, and Theorem 3 guarantees that the platform objective increases with every split. In particular, this means that allowing for one more scale is always better.

The full proof of Theorem 3 is technically involved and can be found in Appendix A.3. Here we provide an outline. Let $p := \beta(\theta_1) < q := \beta(\theta_2)$. The pairwise objective P_k for using β on all $k = 2m$ ratings is

$$P_k(\theta_1, \theta_2 \mid \beta) = K_{2m}(p, q) := \mathbb{E}[\mathrm{sign}(\mathrm{Binom}(2m, q) - \mathrm{Binom}(2m, p))]$$

where $\mathrm{sign}(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x = 0 \\ -1 & \text{if } x < 0 \end{cases}$ is the sign function. To formalize the split, we define the bounding functions $B^+(x) := \min\{2x, 1\}$ and $B^-(x) := \max\{2x - 1, 0\}$ so that $\beta^+ = B^+ \circ \beta$ and $\beta^- = B^- \circ \beta$. An item with quality θ and $p = \beta(\theta)$ has $Y_p \sim \mathrm{Binom}(m, B^+(p)) + \mathrm{Binom}(m, B^-(p))$ positive ratings, hence the pairwise objective for using β^+ and β^- on m ratings each is

$$P_k(\theta_1, \theta_2 \mid \beta^+, \beta^-) = \widetilde{K}_{2m}(p, q) := \mathbb{E}[\mathrm{sign}(Y_q - Y_p)]\,.$$

To prove Theorem 3 it suffices to show that $\widetilde{K}_{2m}(p, q) \geq K_{2m}(p, q)$ for all $p < q$, as the result follows by taking a weighted average over $\theta_1 < \theta_2$ according to the weight function $w(\theta_1, \theta_2)$. In

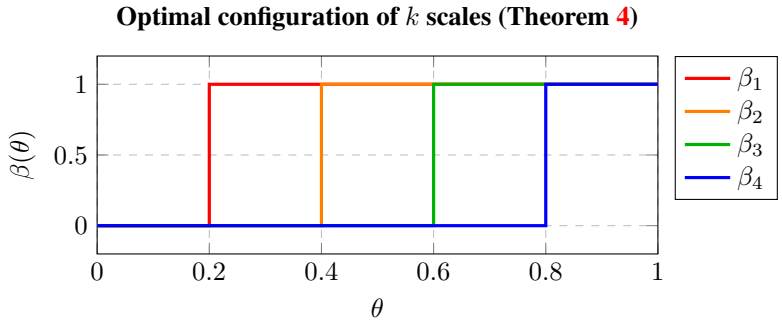


Figure 2: For Kendall's τ function, the optimal k scales for k ratings are 0-1 step functions with evenly spaced thresholds in $[0, 1]$. The figure illustrates the optimal configuration of $k = 4$ scales.

Appendix A.3 we show that $\widetilde{K}_{2m}(p, q) \geq K_{2m}(p, q)$ via a sequence of rating hybrids: we start with the original $k = 2m$ copies of β-scales and replace a pair of β-scales by one β^+-scale and one β^--scale, one pair at a time. The key claim is that each such replacement weakly increases the objective W_k, and strictly increases the pairwise objective for every pair with $\beta(\theta_1) < \beta(\theta_2)$ other than $(\beta(\theta_1), \beta(\theta_2)) = (0, 1)$. This is what makes the gain strict unless β is a 0-1 step function or a constant. A constant scale has objective 0 before and after splitting.

5 Global result: deriving the optimal multiple scales

5.1 One rating per scale

In this section, we derive the optimal configuration of scales when the platform may choose k possibly different rating scales, one for each of the k ratings, so that each scale is used for exactly one rating.

Theorem 4. *Consider any weight function $w(\theta_1, \theta_2)$. Given k scales $\beta_1, \ldots, \beta_k$ to be used for k ratings, where the i-th rating uses scale β_i, the maximum platform objective is achieved by 0-1 step functions $\beta_i^*(\theta) = \mathbf{1}[\theta \geq \tau_i]$ for all $i \in [k]$, with threshold values $0 \leq \tau_1 \leq \cdots \leq \tau_k \leq 1$ that depend on $w(\theta_1, \theta_2)$. The thresholds are distinct and lie in $(0, 1)$ whenever $w > 0$ almost everywhere. When $w(\theta_1, \theta_2)$ is Kendall's τ function, the threshold values are evenly spaced in $[0, 1]$, that is, $\tau_i = \frac{i}{k+1}$ for all $i \in [k]$.*

The main idea in the proof of Theorem 4 is to work with the aggregated outcome of the k ratings, that is, with the aggregation rule of Section 3. For an item of quality θ, let $n(\theta) \in \{0, 1, \ldots, k\}$ denote the total number of positive ratings it receives, and let $p(\theta) = (p_0(\theta), \ldots, p_k(\theta))$ be its distribution in the simplex $\Delta(\{0, 1, \ldots, k\})$, where $p_i(\theta)$ is the probability that the item receives exactly i positive ratings. The platform objective depends on the scales only through the map $\theta \mapsto p(\theta)$, and the proof proceeds in four steps. *Step 1: relaxation.* Because the scales are nondecreasing, the distribution of $n(\theta)$ is stochastically nondecreasing in θ. We relax the problem to arbitrary families of score distributions with this property and represent each family through its quantile functions. *Step 2: deterministic and monotone assignments.* A coupling argument shows that such a family never does better than one of the deterministic nondecreasing assignments $\theta \mapsto n(\theta)$ that it is a mixture of, so the problem reduces to partitioning $[0, 1]$ into $k+1$ ordered quality intervals. *Step 3: interval lengths.* For Kendall's τ function, a Lagrange multiplier computation shows that the intervals should have equal length. This is the only step that depends on the specific weight function. *Step 4: implementation.* Any such partition is realized by 0-1 step functions with thresholds at the interval boundaries, so the relaxation is tight. The full proof is in Appendix A.4, and Steps 3 and 4 are reused for the adaptive setting in Section 6.

Non-uniform weight functions. Since only Step 3 depends on the specific weight function, for any weight function $w(\theta_1, \theta_2)$ the optimal k scales are still a collection of 0-1 step functions, but how the k thresholds are spaced in $[0, 1]$ depends on $w(\theta_1, \theta_2)$: the thresholds minimize the total weight of the pairs of qualities that fall in the same interval, as shown in Eq. (9) of Appendix A.4. Appendix A.8 works out the thresholds for Spearman's ρ function, for which they turn out to be evenly spaced as well, and for a weight function that only values the top half of the items, for which they are not. Thus,

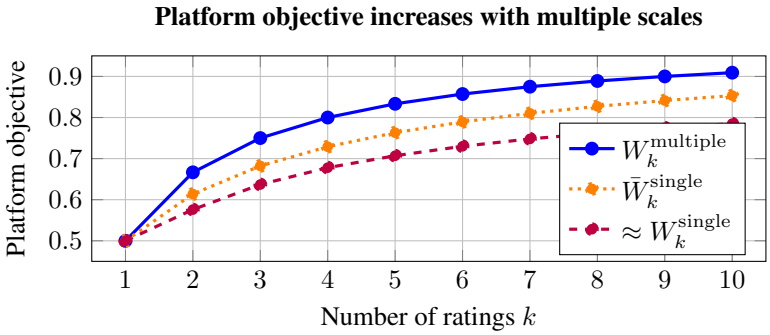


Figure 3: For $k \geq 2$, using a different scale for each rating has a concrete advantage over using the same scale. The approximation for W_k^{single} is calculated based on the best discretized scale computed in Figure 5.

the optimal multiple-scale binary rating system partitions items into $k+1$ ordered quality intervals, so that items in different intervals receive different numbers of positive ratings. We can distinguish items that lie in different intervals with probability 1. Items that lie in the same interval cannot be distinguished because they accumulate the same number of positive ratings.

Optimal objective value for k scales. By Theorem 4, for Kendall's τ function the maximum platform objective is achieved using 0-1 step functions with uniformly spaced thresholds in $[0, 1]$, so

$$W_k^{\text{multiple}} = 2 \int_{\theta_1 < \theta_2} \mathbf{1}\left[\theta_1, \theta_2 \text{ lie in different intervals}\right] d\theta_1 d\theta_2 = 1 - \frac{1}{k+1}$$

since the probability that two items lie in the same quality interval is $\frac{1}{k+1}$.

Comparison with single scale. Figure 3 compares how much larger the maximum platform objective W_k^{multiple} is for multiple scales compared to the maximum objective W_k^{single} for a single scale. Recall from Section 3 that computing W_k^{single} exactly is intractable, so we employed two estimation methods: the first is to upper bound the platform objective by $\bar{W}_k^{\text{single}}$ and the second is to estimate $\approx W_k^{\text{single}}$ by solving for the best discretized scale. For example, at $k = 10$ the best discretized single scale achieves ≈ 0.785, the upper bound is $\bar{W}_{10}^{\text{single}} \approx 0.853$, and ten specialized scales achieve $W_{10}^{\text{multiple}} = \frac{10}{11} \approx 0.909$. Because the multiple-scale value exceeds even the upper bound, the gap between the solid and the dotted curves in Figure 3 is a rigorous lower bound on the advantage of heterogeneous scales, and the gap to the dashed curve is our best estimate of it.

5.2 Multiple ratings per scale

In this section we consider a setting in which we allow multiple rating scales, but not so many as to have one rating per scale. If solving for the best single scale for k ratings is hard, then computing the optimal set of ℓ scales each used for $\frac{k}{\ell}$ ratings must be challenging as well. To estimate the optimal configuration of ℓ scales, we again employ discretization. We split $[0, 1]$ into n equally spaced intervals and solve for the step functions $\beta_1, \ldots, \beta_\ell$ that are constant within each interval and that maximize W_k. Examples of this numerical computation for $\ell \in \{2, 3, 4\}$ are shown in Figure 4.

We make the empirical observation that the optimal configuration of multiple scales exhibits *specialization*: each scale focuses on distinguishing items from a certain region. Furthermore, as the number ℓ of allowed scales increases to k, the rating scales $\beta_1, \ldots, \beta_\ell$ converge to 0-1 step functions within their respective specialized regions. In the limiting case $\ell = k$, Theorem 4 tells us that each β_i is fully specialized: the job of β_i is to exactly distinguish whether an item has quality higher or lower than the i-th threshold.

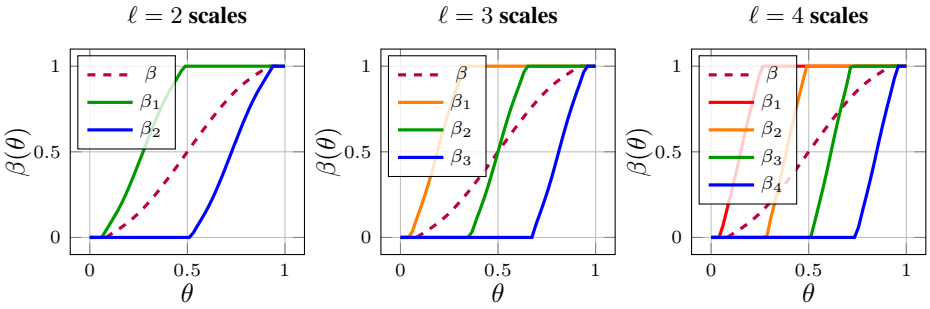


Figure 4: The best discretized sets of ℓ scales for $\ell \in \{2, 3, 4\}$ are graphed alongside the best discretized single scale β. All graphs use Kendall's τ function, $k = 24$ ratings, and a discretization of $n = 50$ steps. For better visualization, we connect the dots between steps instead of graphing the discontinuous step functions themselves.

6 Designing adaptive rating scales

In this section, we consider an adaptive setting where k users arrive sequentially, and the scale shown to each user may depend on the binary ratings of all previous users. Our main result is that adaptivity provides an exponential gain in the maximum platform objective W_k.

Adaptive model. The scale β_j shown to the j-th user may depend on the ratings $(r_1, \dots, r_{j-1}) \in \{0,1\}^{j-1}$ of the first $j-1$ users. Equivalently, the design is a complete binary tree of depth k: a node at level j is labeled by the nondecreasing scale used for the j-th rating, and its two children correspond to the two possible binary responses, which determine the scale used for the next rating. Under scale β_j, an item of quality θ receives a positive rating independently with probability $\beta_j(\theta)$, exactly as before. After k ratings, an item's *outcome* is its full response sequence $\omega = (r_1, \dots, r_k) \in \Omega := \{0,1\}^k$, that is, the leaf of the tree that the item reaches, and the platform ranks items by their outcomes according to a total order $\prec$ on Ω of its choice. The pairwise ranking accuracy between qualities $\theta_1 < \theta_2$ is thus

$$P_k(\theta_1, \theta_2) := \mathbb{P}[\omega(\theta_1) \prec \omega(\theta_2)] - \mathbb{P}[\omega(\theta_1) \succ \omega(\theta_2)]\,,$$

and W_k is obtained from P_k exactly as in Section 3. With $k = 2$, for example, the sequences 01 and 10 tie under the count-based rule of Section 3, but they lead to different leaves of the tree and may be ranked differently here. The adaptive model nevertheless contains the non-adaptive one, since a tree whose scales ignore previous responses, with outcomes ordered by their number of ones and ties allowed, recovers the objective of Section 3 exactly.

The adaptive model therefore changes both which questions are asked and how answers are aggregated: items are ranked by their sequence of responses rather than by their number of positive ratings. The next proposition shows that the second change is necessary, so the exponential gain in Theorem 6 below should be attributed to the combination of adaptive questions and ranking by the full response sequence, not to adaptivity alone. Fully sequential adaptation also requires each rating to be collected after the previous ones. When ratings arrive in parallel, the batch-adaptive objective at the end of this section quantifies the resulting loss.

Proposition 5 (Adaptivity with count-based ranking)**.** *Let w be Kendall's τ function. Suppose the platform elicits ratings adaptively with any tree of scales of depth k, but still ranks items by the number of positive ratings $n_k(\theta) \in \{0, 1, \dots, k\}$. Then $W_k \le 1 - \frac{1}{k+1} = W_k^{\text{multiple}}$, which is already attained by the non-adaptive scales of Theorem 4.*

Proof. Any tree of scales induces, for each quality θ, a random variable $n_k(\theta) \in \{0, 1, \dots, k\}$. Lemma 11 in Appendix A.4 bounds the objective of *every* such collection of random variables, however it is generated, by the objective of the best partition of $[0,1]$ into $k+1$ ordered intervals, which for Kendall's τ function equals $1 - \frac{1}{k+1}$ by Theorem 4. □

In words, with only $k+1$ possible scores the platform can distinguish at most $k+1$ groups of items, whatever questions it asks. Adaptivity creates 2^k distinguishable outcomes, and the platform

must rank by them to benefit. By Lemma 11, the order that the platform uses on the 2^k outcomes is immaterial for the optimal value.

Theorem 6. *Consider any weight function $w(\theta_1, \theta_2)$ that is nonincreasing in θ_1 and nondecreasing in θ_2, such as Kendall's τ function or Spearman's ρ function. Given k scales $\beta_1, \ldots, \beta_k$ to be adaptively chosen for k ratings, where the j-th rating uses scale β_j and can depend on the responses to the first $j-1$ scales, and given that items are ranked by their response sequence $\omega \in \{0,1\}^k$, the maximum platform objective is achieved by partitioning $[0,1]$ into 2^k ordered quality intervals and performing a binary search to determine which of these intervals the item lies in. Concretely, every scale in the tree is a 0-1 step function whose threshold is the middle boundary of the set of intervals still consistent with the responses so far, and the outcomes are ordered lexicographically: $\omega_1 \prec \omega_2$ if and only if $\omega_1 < \omega_2$ when read as binary numbers. When w is Kendall's τ function, the 2^k intervals have equal length.*

Proof sketch. The full proof is in Appendix A.6 and follows the four steps of the proof of Theorem 4. Any tree of scales maps each quality θ to a distribution $p(\theta) \in \Delta(\Omega)$ over outcomes, so we relax the problem and optimize over arbitrary maps $\theta \mapsto p(\theta) \in \Delta(\Omega)$ and arbitrary orders $\prec$ on Ω. Unlike in the non-adaptive setting, this family need not be monotone in θ with respect to the order chosen by the platform, so Step 2 of Theorem 4 is replaced by Lemma 11 in Appendix A.4, which shows, for the weight functions in the statement, that the relaxed optimum is still achieved by a deterministic assignment that partitions $[0,1]$ into 2^k ordered intervals, one per outcome, and that the choice of order is immaterial. The interval lengths are then chosen as in Step 3 of Theorem 4. What is new is the implementation: any partition into 2^k ordered intervals is realized by a depth-k tree of 0-1 step functions that binary searches for the interval containing θ, with outcomes ordered lexicographically, so the relaxation is tight. □

Platform objective under Kendall's τ function. For Kendall's τ function, the same Lagrange multipliers analysis as in Theorem 4 yields that the optimal groups in Theorem 6 are a partition of $[0,1]$ into 2^k equal length intervals. The binary search then uses threshold $\frac{1}{2}$ for the first rating, threshold $\frac{1}{4}$ or $\frac{3}{4}$ for the second rating depending on the first response, and so on, so that the outcome $\omega(\theta)$ is simply the first k binary digits of θ. Since the probability that two items lie in the same quality interval is $\frac{1}{2^k}$, the maximum platform objective under Kendall's τ function for k ratings using k adaptive scales is $W_k^{\text{adaptive}} = 1 - \frac{1}{2^k}$. For $k \geq 2$, this is strictly larger than the optimal objective $W_k^{\text{multiple}} = 1 - \frac{1}{k+1}$ under the non-adaptive model. More specifically, k adaptive ratings have the same distinguishing power as $2^k - 1$ non-adaptive ratings.

Partial adaptivity. Fully sequential adaptation requires each user's scale to be set after all previous ratings of the same item have arrived, which may be infeasible when many ratings arrive in parallel. The two models above are the endpoints of a spectrum. Suppose ratings arrive in k/b batches of size b: the b scales within a batch are set simultaneously but may depend on the responses of all previous batches, and items are ranked by the sequence of per-batch positive counts. We show in Appendix A.7 that the maximum objective under Kendall's τ function is $1 - \frac{1}{(b+1)^{k/b}}$, which interpolates between the non-adaptive value $1 - \frac{1}{k+1}$ at $b = k$ and the fully adaptive value $1 - \frac{1}{2^k}$ at $b = 1$. Even a few rounds of adaptation recover a large fraction of the exponential gain.

7 Conclusion and discussion

When designing rating scales on online platforms, we demonstrate the benefits of using heterogeneous scales on different users. This allows user ratings to specialize in distinguishing certain subsets of items, resulting in overall better accuracy. Multiple scales are simple to implement in practice, for example by randomizing the questions shown to different users.

Implementing rating scales. A complementary question is how a platform can build a real rating system that elicits binary responses according to a given scale β. Garg and Johari [2019] describe and validate one such design. The platform starts with a pool $\mathcal{Y}$ of candidate binary questions that differ in the strictness of their wording and a representative set Θ of items whose qualities are known. It first estimates, for each question $y \in \mathcal{Y}$ and each item $\theta \in \Theta$, the probability $\psi(\theta, y)$

of a positive response, and then chooses a distribution H over $\mathcal{Y}$ that minimizes $\sum_{\theta\in\Theta} \left|\beta(\theta) - \sum_{y\in\mathcal{Y}} \psi(\theta, y)H(y)\right|$. Showing each user a question drawn independently from H then implements a scale close to β. The specialized scales of Theorem 4 are implemented by the same procedure with a different target β_i for each user. A 0-1 step function with threshold τ_i corresponds to concentrating H on the question whose strictness matches τ_i.

Garg and Johari [2019] validate this procedure on Amazon Mechanical Turk, where 200 workers rated the English proficiency and argument coherence of ten short paragraphs adapted from TOEFL writing samples published by the Educational Testing Service. The paragraphs span five expert quality levels $\theta \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$, with two paragraphs per level. Each worker answered threshold questions of the form "Would you characterize the performance of this item as [adjective] or better?" for adjectives ranging from "Abysmal" to "Phenomenal," which form the question pool $\mathcal{Y}$. The fitted mixture tracks the target scale closely except for the lowest-quality paragraphs, where the available questions could not separate items that the experts had ranked differently, so the question pool must be rich where the target scale varies. For a Kendall's τ objective, the design asks most users whether the item is "Fair" or better, while objectives that emphasize top items shift mass to the "Good" question.

Limitations. We discuss the main idealizations in our model and how our results should be interpreted in light of them.

Noisy and imperfectly implemented scales. The optimal scales in Theorems 4 and 6 are 0-1 step functions, which assume that users can perfectly determine whether the quality of an item is above or below a given threshold τ. Real questions induce smooth response functions, for example $\beta_i(\theta) = \sigma(a(\theta - \tau_i))$ for a sigmoid σ and a finite slope a, so our results should be read as design principles rather than literal implementation targets. Two of our findings do not depend on this idealization. First, Theorem 3 applies to *any* scale, however noisy: splitting a smooth, non-constant scale into two specialized smooth scales strictly improves the objective. Second, step functions are only optimal when the number of scales equals the number of ratings. When $\ell < k$ scales are shared among k ratings, the best discretized scales in Figure 4 are themselves smooth and specialized, and they continue to outperform the best single scale. What we do not characterize is how fast the gains in Theorems 4 and 6 degrade as the realized scales depart from step functions, for instance as the slope a decreases. We leave such a sensitivity analysis to future work.

User heterogeneity. In our model, the probability $\beta_i(\theta)$ of a positive rating depends on the item and on the question, but not on the user. In practice, different users respond differently to the same question: one user may consider an item their best experience while another finds it only mildly positive. Our results still apply when users are matched to items at random, because the rating of a random user is then $\mathrm{Bern}(\beta_q(\theta))$, where $\beta_q(\theta)$ is the response probability of question q averaged over the user population. This population scale is what a platform estimates from data [Garg and Johari, 2019], and no calibration of individual users is required. Averaging over users is also one reason why realized scales are smoother than step functions. What our theory does not capture is the information in *which* user provided a rating: scales that depend on the user, for example through observable covariates, could specialize further, and we leave this extension to future work.

User experience and fairness. Heterogeneous scales are implemented by randomizing the questions shown to different users. Since the randomization is independent of the item and of the user, every seller and every user faces the same distribution of questions in expectation, so no item is systematically evaluated under a stricter question than others. A platform may nonetheless wish to limit how different the questions shown to different users can be, but Theorem 3 shows that even two different questions already yield a strict improvement.

Open questions. In the non-adaptive setting we follow prior work and aggregate ratings by their empirical mean, while in the adaptive setting we rank by the full response sequence. In both cases, the aggregation rule is fixed and the scales are optimized against it. It is an open question to derive better aggregation methods, and moreover, to jointly design the rating scales and the aggregation methods. More generally, in classic machine learning paradigms, the data distributions are usually assumed exogenous and fixed. The idea of changing the distribution under which people provide data applies more broadly. For example, it would be interesting to jointly design an online recommendation algorithm and a rating scale, which may itself be adaptive.

References

Arslan Aziz, Hui Li, and Rahul Telang. The consequences of rating inflation on platforms: Evidence from a quasi-experiment. *Information Systems Research*, 34(2):590–608, 2023.

A. P. Dawid and A. M. Skene. Maximum likelihood estimation of observer error-rates using the em algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):20–28, 1979.

Chrysanthos Dellarocas and Charles A. Wood. The sound of silence in online feedback: Estimating trading risks in the presence of reporting bias. *Management Science*, 54(3):460–476, 2008.

S.E. Embretson and S.P. Reise. *Item Response Theory for Psychologists*. Multivariate Applications Series. Taylor & Francis, 2013.

Apostolos Filippas, John J. Horton, and Joseph M. Golden. Reputation inflation. *Marketing Science*, 41(4):733–745, 2022.

Andrey Fradkin, Elena Grewal, and David Holtz. Reciprocity and unveiling in two-sided reputation systems: Evidence from an experiment on airbnb. *Marketing Science*, 40(6):1013–1029, November 2021.

Nikhil Garg and Ramesh Johari. Designing optimal binary rating systems. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1930–1939. PMLR, 2019.

Nikhil Garg and Ramesh Johari. Designing informative rating systems: Evidence from an online labor market. *Manufacturing & Service Operations Management*, 23(3):589–605, 2021.

Dale Griffin and Lyle Brenner. *Perspectives on Probability Judgment Calibration*, chapter 9, pages 177–199. John Wiley & Sons, Ltd, 2004.

Alan J. Klockars and Midori Yamagishi. The influence of labels and positions in rating scales. *Journal of Educational Measurement*, 25(2):85–96, 1988.

R. S. MacKay, R. Kenna, R. J. Low, and S. Parker. Calibration with confidence: a principled method for panel assessment. *Royal Society Open Science*, 4(2):160760, 02 2017.

George A. Miller. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 63(2):81–97, 1956.

Chris Nosko and Steven Tadelis. *The Limits of Reputation in Platform Markets: An Empirical Analysis and Field Experiment*. January 2015.

S. R. Paul. Bayesian methods for calibration of examiners. *British Journal of Mathematical and Statistical Psychology*, 34(2):213–223, 1981.

Joaquim F. Pinto Da Costa and Luís A.C. Roque. Limit distribution for the weighted rank correlation coefficient, rw. *REVSTAT-Statistical Journal*, pages Vol. 4 No. 3 (2006): REVSTAT–Statistical Journal, 2006.

Rachmiani Rachmiani, Nabila Kintan Oktadinna, and Tribowo Rachmat Fauzan. The impact of online reviews and ratings on consumer purchasing decisions on e-commerce platforms. *International Journal of Management Science and Information Technology*, 4(2):504–515, December 2024. ISSN 2776-7388. doi: 10.35870/ijmsit.v4i2.3373. URL http://dx.doi.org/10.35870/ijmsit.v4i2.3373.

Magnus Roos, Jörg Rothe, and Björn Scheuermann. How to calibrate the scores of biased reviewers by quadratic programming. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2011.

Agostino Tarsitano. Comparing the effectiveness of rank correlation statistics. Working Papers 200906, Università della Calabria, Dipartimento di Economia, Statistica e Finanza "Giovanni Anania" - DESF, Apr 2009.

Peter Welinder, Steve Branson, Pietro Perona, and Serge Belongie. The multidimensional wisdom of crowds. In *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc., 2010.

A Appendix

A.1 Discretized solution to the optimal single scale

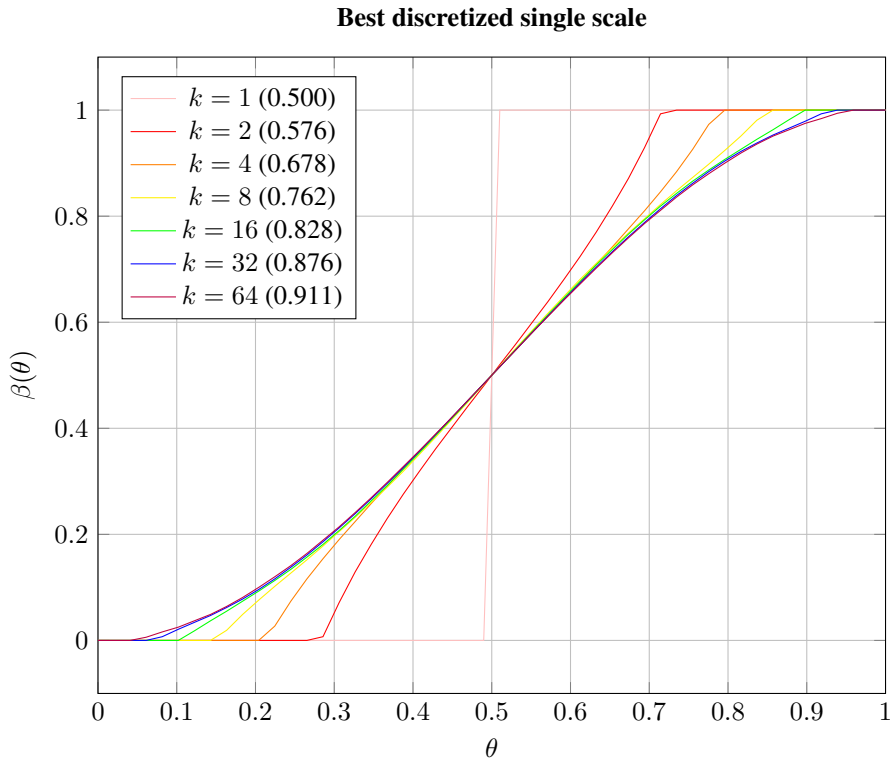


Figure 5: The optimal single scale for Kendall's τ function discretized to $n = 50$ steps is graphed for all numbers $k \in \{1, 2, 4, 8, 16, 32, 64\}$ of ratings. The legend shows the platform objective of each discretized scale. For better visualization, we connect the dots between steps instead of graphing the discontinuous step functions themselves.

A.2 Proof of Proposition 2

Recall from Section 3 that we have two ways to estimate the platform objective for a single scale. The first is by showing a theoretical upper bound $\bar{W}_k^{\text{single}}$, and the second is by solving for the best *discretized* single scale.

To prove the upper bound in Proposition 2, the idea is to derive a *pairwise* bound for every pair of item qualities. Intuitively, for two quality levels $i < j$ with positive-rating probabilities $p_i \le p_j$, if the difference $\delta = p_j - p_i$ is small, then it is hard to correctly rank these two levels using repeated noisy ratings.

Lemma 7. *Let* $0 \le p_i \le p_j \le 1$ *and* $\delta = p_j - p_i$. *Let*

$$S_i := \sum_{r=1}^{k} X_{i,r}, \qquad S_j := \sum_{r=1}^{k} X_{j,r},$$

where $X_{i,r} \overset{\text{iid}}{\sim} \text{Bernoulli}(p_i)$ *and* $X_{j,r} \overset{\text{iid}}{\sim} \text{Bernoulli}(p_j)$. *Then, we have*

$$\mathbb{E}\big[\,\text{sign}(S_j - S_i)\big] \le 1 - (1 - \delta)^k.$$

Proof. To bound the expected sign of $S_j - S_i$, we decompose $S_j - S_i$ into two random variables, one with zero mean and one that captures the bias due to the gap δ. For each round $r \in [k]$, define

$$D_r := X_{j,r} - X_{i,r} \in \{-1, 0, 1\}.$$

Then we have $\mathbb{P}(D_r = 1) = p_j(1 - p_i)$ and $\mathbb{P}(D_r = -1) = p_i(1 - p_j)$. Let $b = p_i(1 - p_j)$. Since $\delta = p_j - p_i$, we have

$$\mathbb{P}(D_r = 1) = \delta + b, \qquad \mathbb{P}(D_r = -1) = b, \qquad \mathbb{P}(D_r = 0) = 1 - \delta - 2b.$$

We first decompose the random variable D_r into a bias term and a symmetric term. We define the random variable $G_r \sim \text{Bernoulli}(\delta)$. Independently of G_r, let $A_r \in \{-1, 0, 1\}$ be a random variable with the symmetric distribution

$$\mathbb{P}(A_r = 1) = \mathbb{P}(A_r = -1) = \frac{b}{1 - \delta}, \qquad \mathbb{P}(A_r = 0) = 1 - \frac{2b}{1 - \delta}.$$

This distribution is well-defined since $0 \le 2b/(1 - \delta) \le 1$ for any $0 \le p_i \le p_j \le 1$. Since $G_r + (1 - G_r)A_r$ has the same distribution as D_r, we may realize

$$D_r = G_r + (1 - G_r)A_r.$$

Hence, summing over all rounds, the difference between the two scores is

$$S_j - S_i = \sum_{r=1}^{k} D_r = G + A,$$

where

$$G := \sum_{r=1}^{k} G_r \sim \text{Binomial}(k, \delta), \qquad A := \sum_{r=1}^{k} (1 - G_r)A_r.$$

Note that conditional on $(G_1, \dots, G_k)$, the random variable A is symmetric about 0, since it is a sum of independent symmetric terms. Therefore, we have

$$\mathbb{E}\big[\operatorname{sgn}(G + A) \mid G_1, \dots, G_k\big] = 0 \quad \text{if } G = 0,$$

while trivially

$$\mathbb{E}\big[\operatorname{sgn}(G + A) \mid G_1, \dots, G_k\big] \le 1 \quad \text{if } G > 0.$$

Thus, we obtain

$$\mathbb{E}\big[\operatorname{sgn}(S_j - S_i)\big] \le \mathbb{P}(G > 0) = 1 - (1 - \delta)^k.$$

□

We now use Lemma 7 to prove Proposition 2.

Proof of Proposition 2. Fix any nondecreasing scale β and discretize it to n levels $p_i := \beta(i/n)$, so that

$$0 \le p_1 \le p_2 \le \cdots \le p_n \le 1.$$

By Lemma 7, for every pair $i < j$, we have $\mathbb{E}\big[\operatorname{sgn}(S_j - S_i)\big] \le \phi_k(p_j - p_i)$, where $\phi_k(x) := 1 - (1 - x)^k$.

Therefore, the platform's objective on the discretized scale is

$$\frac{1}{\binom{n}{2}} \sum_{1 \le i < j \le n} \mathbb{E}\big[\operatorname{sgn}(S_j - S_i)\big] \le \frac{1}{\binom{n}{2}} \sum_{1 \le i < j \le n} \phi_k(p_j - p_i).$$

We group pairs according to their index gap $\ell = j - i$. Then

$$\sum_{1 \le i < j \le n} \phi_k(p_j - p_i) = \sum_{\ell=1}^{n-1} \sum_{i=1}^{n-\ell} \phi_k(p_{i+\ell} - p_i).$$

Since ϕ_k is increasing and concave on $[0, 1]$, by Jensen's inequality, we have

$$\sum_{i=1}^{n-\ell} \phi_k(p_{i+\ell} - p_i) \le (n - \ell)\, \phi_k\left(\frac{1}{n - \ell} \sum_{i=1}^{n-\ell} (p_{i+\ell} - p_i)\right).$$

Moreover, because $0 \le p_i \le 1$ for all i, when $\ell \le n/2$, so that $n - \ell \ge \ell$, we have

$$\sum_{i=1}^{n-\ell}(p_{i+\ell} - p_i) = \sum_{i=n-\ell+1}^{n} p_i - \sum_{i=1}^{\ell} p_i \le \ell,$$

and when $\ell > n/2$, so that $n - \ell < \ell$, we have

$$\sum_{i=1}^{n-\ell}(p_{i+\ell} - p_i) = \sum_{i=\ell+1}^{n} p_i - \sum_{i=1}^{n-\ell} p_i \le n - \ell.$$

Thus, we have

$$\frac{1}{n-\ell}\sum_{i=1}^{n-\ell}(p_{i+\ell} - p_i) \le \frac{\min\{\ell, n-\ell\}}{n-\ell}.$$

Hence, we obtain

$$\sum_{i=1}^{n-\ell}\phi_k(p_{i+\ell} - p_i) \le (n-\ell)\,\phi_k\left(\min\left\{1, \frac{\ell}{n-\ell}\right\}\right).$$

Therefore, our objective is at most

$$U_{n,k} := \frac{1}{\binom{n}{2}}\sum_{\ell=1}^{n-1}(n-\ell)\left[1 - \left(1 - \min\left\{1, \frac{\ell}{n-\ell}\right\}\right)^k\right].$$

Since any scale is approximated arbitrarily well by its discretizations, taking $n \to \infty$ gives the bound for the original scale, where

$$\lim_{n\to\infty} U_{n,k} = 2\int_0^1 (1-x)\left[1 - \left(1 - \min\left\{1, \frac{x}{1-x}\right\}\right)^k\right] dx.$$

For $x \le 1/2$,

$$1 - \frac{x}{1-x} = \frac{1-2x}{1-x},$$

while for $x \ge 1/2$, the minimum term equals 1. Hence

$$\begin{aligned}\lim_{n\to\infty} U_{n,k} &= 2\int_0^{1/2}(1-x)\left[1 - \left(\frac{1-2x}{1-x}\right)^k\right] dx + 2\int_{1/2}^1 (1-x)\,dx \\ &= 1 - 2\int_0^{1/2}\frac{(1-2x)^k}{(1-x)^{k-1}}\,dx.\end{aligned}$$

Finally, using the change of variables $y = 1 - 2x$, we get

$$2\int_0^{1/2}\frac{(1-2x)^k}{(1-x)^{k-1}}\,dx = 2^{k-1}\int_0^1 \frac{y^k}{(1+y)^{k-1}}\,dy.$$

Therefore, our objective is upper bounded by

$$\bar{W}_k^{\text{single}} = 1 - 2^{k-1}\int_0^1 \frac{y^k}{(1+y)^{k-1}}\,dy.$$

Finally, we note that this bound is strictly smaller than $W_k^{\text{multiple}} = 1 - \frac{1}{k+1}$ for every $k \ge 2$, since

$$\begin{aligned}2^{k-1}\int_0^1 \frac{y^k}{(1+y)^{k-1}}\,\mathrm{d}y &= \int_0^1 \left(\frac{2}{1+y}\right)^{k-1} y^k\,\mathrm{d}y \\ &> \int_0^1 y^k\,\mathrm{d}y = \frac{1}{k+1},\end{aligned}$$

where the inequality uses $2/(1+y) > 1$ for $y < 1$. This completes the proof. □

A.3 Proof of Theorem 3

As discussed in Section 4, to prove Theorem 3 it suffices to show that $\widetilde{K}_{2m}(p,q) \geq K_{2m}(p,q)$ for all $p < q$. We do so via a sequence of rating hybrids: we start with the original $k = 2m$ copies of β-scales and replace a pair of β-scales by one β^+-scale and one β^--scale, one pair at a time. We claim that every such replacement weakly increases the objective W_k. Define

$$a := q(1-p), \quad c := p(1-q), \quad r := \frac{a}{c} = \frac{q(1-p)}{p(1-q)} > 1,$$

$$U := \mathrm{Bern}(q) - \mathrm{Bern}(p)\,,$$
$$W := \mathrm{Bern}\big(B^+(q)\big) + \mathrm{Bern}\big(B^-(q)\big) - \mathrm{Bern}\big(B^+(p)\big) - \mathrm{Bern}\big(B^-(p)\big)\,,$$

where all Bernoulli random variables are independent. Here U is the difference between the ratings of the two items under one copy of the original scale β, and W is the difference between their total ratings under one β^+-scale and one β^--scale. Note that $\mathbb{E}[U] = \mathbb{E}[W] = q - p$: the split preserves the expected number of positive ratings, and the gain comes from the change in the distribution around this mean.

The ratio r is well defined when $c > 0$, that is, when $0 < p$ and $q < 1$. In the boundary cases $p = 0$ or $q = 1$ we have $c = 0$, and then every negative term in U and W vanishes. Hence $U \geq 0$ and $W \geq 0$ almost surely, so that $\mathbb{P}[U = -t] = \mathbb{P}[W = -t] = 0$ for all $t \geq 1$. In these cases the ratio bounds below hold trivially with the convention $r = \infty$, every sum below that involves $\mathbb{P}[D = -t]$ for $t \geq 1$ vanishes, and the proofs simplify accordingly. We therefore assume $0 < p < q < 1$ from now on and leave the routine adjustments to the reader.

We introduce three helpful technical lemmas.

Lemma 8 (One pair bias)**.** *For every integer $t \geq 1$,*

$$\mathbb{P}[U = t] \geq r^t \cdot \mathbb{P}[U = -t]\,, \qquad \mathbb{P}[W = t] \geq r^t \cdot \mathbb{P}[W = -t]\,.$$

Proof. The claim for U is immediate: $U \in \{-1, 0, 1\}$ and

$$\mathbb{P}[U = 1] = q(1-p) = a = r \cdot c = r \cdot \mathbb{P}[U = -1]\,.$$

For $t \geq 2$, both $\mathbb{P}[U = t]$ and $\mathbb{P}[U = -t]$ are zero.

For W, we distinguish three cases.

Case 1: $0 \leq p < q \leq \frac{1}{2}$**.** Then $B^-(p) = B^-(q) = 0$, $B^+(p) = 2p$, and $B^+(q) = 2q$. Hence we have

$$W = \gamma_q - \gamma_p, \qquad \gamma_q \sim \mathrm{Bern}(2q)\,,\ \gamma_p \sim \mathrm{Bern}(2p)\,.$$

Hence $W \in \{-1, 0, 1\}$ and

$$\frac{\mathbb{P}[W = 1]}{\mathbb{P}[W = -1]} = \frac{2q(1-2p)}{2p(1-2q)} \geq \frac{q(1-p)}{p(1-q)} = r,$$

where the inequality is from $(1-2p)(1-q) - (1-p)(1-2q) = q - p \geq 0$. If $q = \frac{1}{2}$, then $\mathbb{P}[W = -1] = 0$ and there is nothing to prove. Thus $\mathbb{P}[W = 1] \geq r \cdot \mathbb{P}[W = -1]$. Since $\mathbb{P}[W = \pm t] = 0$, the bound also holds for $t \geq 2$.

Case 2: $\frac{1}{2} \leq p < q \leq 1$**.** Then $B^+(p) = B^+(q) = 1$ and

$$B^-(p) = 2p - 1, \qquad B^-(q) = 2q - 1.$$

So

$$W = \eta_q - \eta_p, \qquad \eta_q \sim \mathrm{Bern}(2q-1)\,,\ \eta_p \sim \mathrm{Bern}(2p-1)\,.$$

Hence $W \in \{-1, 0, 1\}$ and

$$\frac{\mathbb{P}[W = 1]}{\mathbb{P}[W = -1]} = \frac{(2q-1)(1-p)}{(2p-1)(1-q)} \geq \frac{q(1-p)}{p(1-q)} = r,$$

where the inequality is from $p(2q-1) - q(2p-1) = q - p \geq 0$. If $p = \frac{1}{2}$, then $\mathbb{P}[W = -1] = 0$ and there is nothing to prove. Again $\mathbb{P}[W = \pm t] = 0$ for all $t \geq 2$, so the bound holds for all $t \geq 1$.

Case 3: $0 \le p \le \frac{1}{2} \le q \le 1$**.** Then

$$B^+(p) = 2p, \qquad B^-(p) = 0, \qquad B^+(q) = 1, \qquad B^-(q) = 2q - 1.$$

So

$$W = 1 + \eta_q - \gamma_p \in \{0, 1, 2\}.$$

Therefore $\mathbb{P}[W = -1] = \mathbb{P}[W = -2] = 0$, and the desired inequalities are immediate. □

Lemma 9 (Multi-pair bias)**.** *Let $X_1, \dots, X_n$ be independent integer-valued random variables such that $\mathbb{P}[X_i = t] \ge r^t \cdot \mathbb{P}[X_i = -t]$ for all i and $t \ge 1$. Then their sum $D := X_1 + \cdots + X_n$ satisfies $\mathbb{P}[D = t] \ge r^t \cdot \mathbb{P}[D = -t]$ for all $t \ge 1$.*

Proof. Fix $t \ge 1$. For every integer vector $(x_1, \dots, x_n)$ with

$$x_1 + \cdots + x_n = t,$$

the reflected vector $(-x_1, \dots, -x_n)$ has sum $-t$, and

$$\prod_{i=1}^{n} \mathbb{P}[X_i = x_i] \ge \prod_{i=1}^{n} r^{x_i} \mathbb{P}[X_i = -x_i] = r^t \prod_{i=1}^{n} \mathbb{P}[X_i = -x_i]\,.$$

Summing over all vectors with total t gives

$$\mathbb{P}[D = t] \ge r^t \cdot \mathbb{P}[D = -t]\,.$$

□

Lemma 10 (One replacement step)**.** *Let U_1 and U_2 be independent copies of U, and let D be an integer-valued random variable such that D, U_1, U_2 and W are mutually independent. Assume that*

$$\mathbb{P}[D = t] \ge r^t \cdot \mathbb{P}[D = -t] \qquad \textit{for every integer } t \ge 1.$$

Then

$$\mathbb{E}[\operatorname{sign}(D + W)] \ge \mathbb{E}[\operatorname{sign}(D + U_1 + U_2)]\,.$$

Proof. Set

$$Z := U_1 + U_2, \qquad \delta_d := \mathbb{E}[\operatorname{sign}(d + W)] - \mathbb{E}[\operatorname{sign}(d + Z)]\,.$$

Then

$$\mathbb{E}[\operatorname{sign}(D + W)] - \mathbb{E}[\operatorname{sign}(D + Z)] = \sum_{d \in \mathbb{Z}} \mathbb{P}[D = d] \cdot \delta_d.$$

Intuitively, because the split ratings W have lower variance than $U_1 + U_2$ and satisfy the ratio bound, replacing $U_1 + U_2$ by W shifts probability mass from negative to positive outcomes. We consider two cases.

Case 1: $0 \le p < q \le \frac{1}{2}$ **or** $\frac{1}{2} \le p < q \le 1$**.** Since $|W|, |Z|$ are bounded by 2, $d + W$ and $d + Z$ have the same sign as d for $|d| \ge 3$, namely $\delta_d = 0$ for $|d| \ge 3$. Hence, we consider $d \in \{\pm 2, \pm 1, 0\}$. Note that

$$\delta_{-2} = -a^2, \qquad \delta_0 = a^2 - c^2, \qquad \delta_2 = c^2,$$

while

$$(\delta_{-1}, \delta_1) = (-L, L),$$

where

$$L = \begin{cases} 2pq(p + q - pq), & 0 < p < q \le \frac{1}{2}, \\ 2(1-p)(1-q)(1-pq), & \frac{1}{2} \le p < q < 1. \end{cases}$$

In particular, $L \ge 0$. Therefore

$$\begin{aligned} \mathbb{E}[\operatorname{sign}(D + W)] - \mathbb{E}[\operatorname{sign}(D + Z)] = \mathbb{P}[D = 0]\,(a^2 - c^2) + L\big(\mathbb{P}[D = 1] - \mathbb{P}[D = -1]\big) \\ + c^2 \mathbb{P}[D = 2] - a^2 \mathbb{P}[D = -2]\,. \end{aligned}$$

By assumption,

$$\mathbb{P}[D = 1] \ge \mathbb{P}[D = -1]\,, \qquad \mathbb{P}[D = 2] \ge r^2 \cdot \mathbb{P}[D = -2]\,.$$

Since $a = rc$, we have

$$c^2\mathbb{P}[D=2] - a^2\mathbb{P}[D=-2] = c^2\big(\mathbb{P}[D=2] - r^2\mathbb{P}[D=-2]\big) \geq 0.$$

Also $a > c$, so $a^2 - c^2 \geq 0$. Hence every term is nonnegative, and

$$\mathbb{E}[\mathrm{sign}(D+W)] \geq \mathbb{E}[\mathrm{sign}(D+Z)]\,.$$

Case 2: $0 < p \leq \frac{1}{2} \leq q < 1$**.** Again $\delta_d = 0$ for $|d| \geq 3$, and a direct calculation gives

$$\delta_1 = 2p(1-q)(1-q+pq), \qquad \delta_{-1} = -\delta_1, \qquad \delta_2 = c^2,$$
$$\delta_0 = (p+q-1)^2 + 2p(q-p)(1-q) \geq 0,$$

and

$$r\delta_1 + \delta_{-1} = 2(q-p)(1-q+pq) \geq 0, \qquad r^2\delta_2 + \delta_{-2} = (1-2p)(2q-1) \geq 0.$$

Therefore

$$\begin{aligned}\mathbb{E}[\mathrm{sign}(D+W)] - \mathbb{E}[\mathrm{sign}(D+Z)] = \mathbb{P}[D=0]\,\delta_0 + \mathbb{P}[D=1]\,\delta_1 + \mathbb{P}[D=-1]\,\delta_{-1} \\ + \mathbb{P}[D=2]\,\delta_2 + \mathbb{P}[D=-2]\,\delta_{-2}.\end{aligned}$$

Since $\delta_1, \delta_2 \geq 0$ and D satisfies the ratio bound,

$$\mathbb{P}[D=1]\,\delta_1 + \mathbb{P}[D=-1]\,\delta_{-1} \geq \mathbb{P}[D=-1]\,\big(r\delta_1 + \delta_{-1}\big) \geq 0,$$
$$\mathbb{P}[D=2]\,\delta_2 + \mathbb{P}[D=-2]\,\delta_{-2} \geq \mathbb{P}[D=-2]\,\big(r^2\delta_2 + \delta_{-2}\big) \geq 0.$$

Together with $\mathbb{P}[D=0]\,\delta_0 \geq 0$, this shows that

$$\mathbb{E}[\mathrm{sign}(D+W)] \geq \mathbb{E}[\mathrm{sign}(D+Z)]\,.$$

□

Now we use the above three lemmas to prove that $\widetilde{K}_{2m}(p,q) \geq K_{2m}(p,q)$ for $p < q$, which implies Theorem 3.

Proof of Theorem 3. For the fixed pair (θ_1, θ_2), take independent copies $U_1, \ldots, U_{2m} \overset{\text{i.i.d.}}{\sim} U$ and $W_1, \ldots, W_m \overset{\text{i.i.d.}}{\sim} W$. Then

$$K_{2m}(p,q) = \mathbb{E}[\mathrm{sign}(U_1 + \cdots + U_{2m})]\,, \qquad \widetilde{K}_{2m}(p,q) = \mathbb{E}[\mathrm{sign}(W_1 + \cdots + W_m)]\,.$$

For $j = 0, 1, \ldots, m$, define

$$S_j := W_1 + \cdots + W_j + U_{2j+1} + \cdots + U_{2m}.$$

Thus

$$S_0 = U_1 + \cdots + U_{2m}, \qquad S_m = W_1 + \cdots + W_m,$$

and so

$$K_{2m}(p,q) = \mathbb{E}[\mathrm{sign}(S_0)]\,, \qquad \widetilde{K}_{2m}(p,q) = \mathbb{E}[\mathrm{sign}(S_m)]\,.$$

Now fix $j \in \{0, \ldots, m-1\}$, and write

$$D_j := W_1 + \cdots + W_j + U_{2j+3} + \cdots + U_{2m}.$$

Then

$$S_j = D_j + U_{2j+1} + U_{2j+2}, \qquad S_{j+1} = D_j + W_{j+1}.$$

By Lemma 8, every summand of D_j satisfies the ratio bound

$$\mathbb{P}[X=t] \geq r^t \cdot \mathbb{P}[X=-t] \qquad (t \geq 1).$$

Hence Lemma 9 implies that D_j also satisfies

$$\mathbb{P}[D_j=t] \geq r^t \cdot \mathbb{P}[D_j=-t] \qquad (t \geq 1).$$

Applying Lemma 10 to D_j, we obtain

$$\mathbb{E}[\operatorname{sign}(S_{j+1})] \geq \mathbb{E}[\operatorname{sign}(S_j)]\,.$$

Since this holds for all $j = 0, \ldots, m-1$, iterating yields

$$\widetilde{K}_{2m}(p,q) = \mathbb{E}[\operatorname{sign}(S_m)] \geq \mathbb{E}[\operatorname{sign}(S_0)] = K_{2m}(p,q),$$

as desired.

Strictness. The first replacement step, with $j = 0$, already yields a strict gain whenever $0 \leq p < q \leq 1$ and $(p,q) \neq (0,1)$. Indeed, in the proof of Lemma 10 every summand is nonnegative, and the summand for $d = 0$ equals $\mathbb{P}[D_0 = 0]\ \delta_0$. Here $D_0 = U_3 + \cdots + U_{2m}$ satisfies

$$\mathbb{P}[D_0 = 0] \geq \mathbb{P}[U = 0]^{2m-2} = \big((1-p)(1-q) + pq\big)^{2m-2} > 0$$

unless $(p,q) = (0,1)$, and

$$\delta_0 = a^2 - c^2 = (q-p)\big(q(1-p) + p(1-q)\big) > 0 \qquad \text{in Case 1,}$$

$$\delta_0 = (p+q-1)^2 + 2p(q-p)(1-q) > 0 \qquad \text{in Case 2,}$$

again unless $(p,q) = (0,1)$. In Case 2, $\delta_0 = 0$ forces $p + q = 1$ and $p(q-p)(1-q) = 0$, hence $p = 0$ and $q = 1$. Therefore $\widetilde{K}_{2m}(p,q) > K_{2m}(p,q)$ for all such pairs. On the other hand, $\widetilde{K}_{2m}(p,q) = K_{2m}(p,q)$ when $p = q$, since both sides vanish by symmetry, and when $(p,q) = (0,1)$, since both sides equal 1. Integrating against $w(\theta_1, \theta_2)$, the increase in W_k is strict if and only if $w > 0$ on a set of positive measure of pairs $\theta_1 < \theta_2$ with $\beta(\theta_1) < \beta(\theta_2)$ and $(\beta(\theta_1), \beta(\theta_2)) \neq (0,1)$. If $w > 0$ almost everywhere, this fails only in two cases. Either β is almost everywhere constant, so that every pair has $\beta(\theta_1) = \beta(\theta_2)$, or β is almost everywhere $\{0,1\}$-valued, that is, a 0-1 step function. This proves the last assertion of Theorem 3. □

A.4 Proof of Theorem 4

We give the full proof of Theorem 4. Throughout this section, fix the number of ratings k and a weight function w, which by Section 3 is nonnegative and integrable on $\{\theta_1 < \theta_2\}$. We write $\operatorname{sign}(x) \in \{-1, 0, 1\}$ for the sign function. For a pair of qualities $\theta_1 < \theta_2$ with scores $n(\theta_1)$ and $n(\theta_2)$, the ranking accuracy is then $\operatorname{sign}\,(n(\theta_2) - n(\theta_1))$.

Setup. Any k scales $\beta_1, \ldots, \beta_k$ induce, for each quality θ, a random score

$$n(\theta) = \sum_{i=1}^{k} \operatorname{Bern}(\beta_i(\theta)) \in \{0, 1, \ldots, k\}.$$

We denote its distribution by $p(\theta) = (p_0(\theta), \ldots, p_k(\theta)) \in \Delta(\{0, 1, \ldots, k\})$. Since the ratings of two different items are independent, the platform objective depends on the scales only through the map $\theta \mapsto p(\theta)$:

$$W_k(p) = 2 \int_{\theta_1 < \theta_2} w(\theta_1, \theta_2) \sum_{i,j=0}^{k} \operatorname{sign}(j-i)\, p_i(\theta_1)\, p_j(\theta_2)\, d\theta_1\, d\theta_2. \tag{2}$$

The scales are nondecreasing, and this has an important consequence for the family $\{p(\theta)\}_{\theta \in [0,1]}$: it is *stochastically nondecreasing* in θ. This means that for all $\theta < \theta'$ and all j,

$$F_\theta(j) := \mathbb{P}[n(\theta) \leq j] \ \geq\ F_{\theta'}(j). \tag{3}$$

To see this, draw independent uniform random variables $U_1, \ldots, U_k$ on $[0,1]$ and set $X_i(\theta) := \mathbf{1}\,[U_i \leq \beta_i(\theta)]$. Then $X_i(\theta) \sim \operatorname{Bern}(\beta_i(\theta))$, so for every θ the sum

$$\sum_{i=1}^{k} X_i(\theta)$$

has the law of $n(\theta)$. Since each β_i is nondecreasing, $X_i(\theta) \leq X_i(\theta')$ for all i whenever $\theta < \theta'$. Therefore

$$\sum_{i=1}^{k} X_i(\theta) \leq \sum_{i=1}^{k} X_i(\theta') \quad \text{almost surely,}$$

which implies Eq. (3).

The proof has four steps. Step 1 relaxes the problem to arbitrary stochastically nondecreasing families and represents each such family through its quantile functions. Step 2 shows that such a family never does better than one of the *deterministic* and *monotone* assignments that it is a mixture of, so that the problem reduces to partitioning $[0,1]$ into $k+1$ ordered intervals. Step 3 chooses the lengths of these intervals, and Step 4 shows that any such partition is implemented by 0-1 step functions. Steps 3 and 4 are reused in Appendix A.6 for the adaptive setting. The families arising there need not be stochastically monotone, so Lemma 11 at the end of this section provides a substitute for Steps 1 and 2 that applies to arbitrary families, at the price of an additional assumption on w.

Step 1: relaxation and quantile representation. We relax the problem and maximize Eq. (2) over all measurable maps $\theta \mapsto p(\theta) \in \Delta(\{0,1,\dots,k\})$ that satisfy Eq. (3), not only those induced by k nondecreasing scales. The relaxed optimum upper bounds the objective of every design. Fix such a map p and define its *quantile functions*

$$Q_\theta(u) := \min\{j \in \{0,1,\dots,k\} : F_\theta(j) \geq u\}, \qquad u \in (0,1).$$

We use three standard facts. First, if U is uniform on $(0,1)$, then $Q_\theta(U)$ has the law $p(\theta)$, because

$$\mathbb{P}[Q_\theta(U) \leq j] = \mathbb{P}[U \leq F_\theta(j)] = F_\theta(j).$$

Second, $Q_\theta(u)$ is nondecreasing in u. Third, and most importantly, Eq. (3) is equivalent to

$$Q_\theta(u) \leq Q_{\theta'}(u) \qquad \text{for all } u \in (0,1) \text{ and all } \theta < \theta'. \tag{4}$$

Indeed, $F_\theta \geq F_{\theta'}$ pointwise means that every level j that is admissible for θ' in the minimum defining $Q_{\theta'}(u)$ is also admissible for θ. We also note that the map $(\theta,u) \mapsto Q_\theta(u)$ is jointly measurable, because $\{Q_\theta(u) \leq j\} = \{F_\theta(j) \geq u\}$ and $\theta \mapsto F_\theta(j)$ is measurable.

Now let U and U' be independent uniform random variables on $(0,1)$. For $\theta_1 < \theta_2$, the scores $n(\theta_1)$ and $n(\theta_2)$ are independent with laws $p(\theta_1)$ and $p(\theta_2)$. The pair $(Q_{\theta_1}(U), Q_{\theta_2}(U'))$ has the same joint law, so Eq. (2) can be written as

$$\begin{aligned} W_k(p) &= 2\int_{\theta_1<\theta_2} w(\theta_1,\theta_2)\,\mathbb{E}[\operatorname{sign}(Q_{\theta_2}(U') - Q_{\theta_1}(U))]\,d\theta_1\,d\theta_2 \\ &= \int_0^1\int_0^1 \Psi(u,u')\,du\,du', \end{aligned} \tag{5}$$

where

$$\Psi(u,u') := 2\int_{\theta_1<\theta_2} w(\theta_1,\theta_2)\,\operatorname{sign}(Q_{\theta_2}(u') - Q_{\theta_1}(u))\,d\theta_1\,d\theta_2.$$

The exchange of the expectation and the integral is justified by Fubini's theorem, since the integrand is bounded in absolute value by the integrable function w.

Step 2: an optimal family is deterministic and monotone. For each $u \in (0,1)$, consider the *deterministic* assignment $n^{(u)} : [0,1] \to \{0,1,\dots,k\}$ defined by $n^{(u)}(\theta) := Q_\theta(u)$. For a deterministic assignment n, write its objective as

$$V(n) := 2\int_{\theta_1<\theta_2} w(\theta_1,\theta_2)\,\operatorname{sign}(n(\theta_2) - n(\theta_1))\,d\theta_1\,d\theta_2.$$

Then $V(n) = W_k(\delta_n)$, where δ_n denotes the family of point masses $\theta \mapsto \delta_{n(\theta)}$, and $\Psi(u,u) = V(n^{(u)})$. By Eq. (4), every $n^{(u)}$ is *nondecreasing* in θ. We claim that for all $u,u' \in (0,1)$,

$$\Psi(u,u') + \Psi(u',u) \;\leq\; V\left(n^{(u)}\right) + V\left(n^{(u')}\right). \tag{6}$$

Since $w \geq 0$, it suffices to prove the corresponding pointwise inequality for the integrands. That is, for every $\theta_1 < \theta_2$,

$$\operatorname{sign}(b'-a) + \operatorname{sign}(b-a') \;\leq\; \operatorname{sign}(b-a) + \operatorname{sign}(b'-a'), \tag{7}$$

where $a := Q_{\theta_1}(u)$, $b := Q_{\theta_2}(u)$, $a' := Q_{\theta_1}(u')$ and $b' := Q_{\theta_2}(u')$. By symmetry we may assume $u \leq u'$. Then $a \leq a'$ and $b \leq b'$ because quantile functions are nondecreasing in u, and $a \leq b$ and $a' \leq b'$ by Eq. (4). In particular, the right side of Eq. (7) is a sum of two terms in $\{0,1\}$. We distinguish three cases according to the second term on the left.

- If $b > a'$, then the left side is at most 2. Since $b > a' \geq a$ and $b' \geq b > a'$, the right side equals $1 + 1 = 2$.
- If $b = a'$, then the left side is at most 1. If the right side were 0, we would have $a = b$ and $a' = b'$, hence $a = b = a' = b'$, and the left side would be 0 as well. Otherwise the right side is at least 1.
- If $b < a'$, then $\operatorname{sign}(b - a') = -1$ and the left side is at most 0, while the right side is at least 0.

This proves Eq. (7), and integrating against $w(\theta_1, \theta_2) \geq 0$ proves Eq. (6). Averaging Eq. (6) over independent uniform u, u' and using Eq. (5) gives

$$\begin{aligned} W_k(p) &= \frac{1}{2} \int_0^1 \int_0^1 \left(\Psi(u, u') + \Psi(u', u) \right) du \, du' \\ &\leq \frac{1}{2} \int_0^1 \int_0^1 \left(V(n^{(u)}) + V(n^{(u')}) \right) du \, du' \\ &= \int_0^1 V\left(n^{(u)}\right) du \\ &\leq \sup_{u \in (0,1)} V\left(n^{(u)}\right). \end{aligned}$$

Every $n^{(u)}$ is a nondecreasing map from $[0, 1]$ to $\{0, 1, \ldots, k\}$. In other words, it partitions $[0, 1]$ into $k + 1$ ordered intervals $I_0, I_1, \ldots, I_k$, listed from left to right, where $I_j = \{\theta : n^{(u)}(\theta) = j\}$ and some intervals may be empty. We have therefore shown that

$$\sup_{p \text{ satisfying } Eq.\ (3)} W_k(p) \;=\; \sup_{I_0, \ldots, I_k} W_k(I_0, \ldots, I_k), \tag{8}$$

where the right side ranges over partitions of $[0, 1]$ into $k + 1$ ordered intervals. The inequality "$\geq$" holds because the family of point masses of such a partition satisfies Eq. (3). Two qualities in different intervals are ranked correctly and two qualities in the same interval tie, so the objective of a partition is

$$W_k(I_0, \ldots, I_k) = \bar{w} - 2 \sum_{j=0}^{k} \int_{(I_j \times I_j) \cap \{\theta_1 < \theta_2\}} w(\theta_1, \theta_2) \, d\theta_1 \, d\theta_2, \tag{9}$$

where

$$\bar{w} := 2 \int_{\theta_1 < \theta_2} w(\theta_1, \theta_2) \, d\theta_1 \, d\theta_2$$

is the total weight. Thus the optimal partition minimizes the total weight of pairs that fall in the same interval. The right side of Eq. (8) is the maximum of a continuous function of the boundaries $0 \leq t_1 \leq \cdots \leq t_k \leq 1$ over a compact set, where continuity follows from the absolute continuity of the integral of w. Hence the supremum is attained.

Step 3: the optimal interval lengths. By Eqs. (8) and (9), it remains to choose the lengths $x_0, \ldots, x_k \geq 0$ of the ordered intervals, subject to

$$\sum_{j=0}^{k} x_j = 1.$$

Only this step depends on the specific weight function. For Kendall's τ function, that is, $w \equiv 1$, the weight of the pairs inside an interval of length x is $x^2/2$ and $\bar{w} = 1$. Then Eq. (9) reads

$$W_k(I_0, \ldots, I_k) = 1 - \sum_{j=0}^{k} x_j^2 = 2 \sum_{0 \leq i < j \leq k} x_i x_j.$$

We maximize

$$f(x_0, \ldots, x_k) := \sum_{0 \leq i < j \leq k} x_i x_j$$

subject to the constraint above. By Lagrange multipliers, at an interior optimum the partial derivatives

$$\frac{\partial f}{\partial x_j} = \sum_{i \neq j} x_i = 1 - x_j$$

are all equal, which forces $x_0 = x_1 = \cdots = x_k = \frac{1}{k+1}$. No boundary point, at which some $x_j = 0$, can be optimal. Indeed, moving a length ε from the longest interval, of length $x_{\max}$, into an empty one changes $\sum_j x_j^2$ by

$$(x_{\max} - \varepsilon)^2 + \varepsilon^2 - x_{\max}^2 = -2\varepsilon(x_{\max} - \varepsilon) < 0$$

for small $\varepsilon > 0$, which increases f. Equivalently, the Cauchy–Schwarz inequality gives

$$\sum_{j=0}^{k} x_j^2 \geq \frac{1}{k+1},$$

with equality if and only if all x_j are equal. The optimal intervals are therefore $I_j = [\frac{j}{k+1}, \frac{j+1}{k+1})$, with boundaries $\tau_i = \frac{i}{k+1}$ for $i \in [k]$, and the optimal value is $W_k^{\text{multiple}} = 1 - \frac{1}{k+1}$.

For a general weight function, the optimal boundaries $0 \leq \tau_1 \leq \cdots \leq \tau_k \leq 1$ are the minimizers of the within-interval weight in Eq. (9). Appendix A.8 works out two examples. When $w > 0$ almost everywhere, the optimal intervals are all nonempty, and hence $0 < \tau_1 < \cdots < \tau_k < 1$. Indeed, if some interval were empty, splitting a nonempty interval into two would strictly increase Eq. (9), because the pairs straddling the new boundary would then be ranked correctly instead of tying.

Step 4: implementation by 0-1 step functions. Finally, we show that every partition of $[0, 1]$ into ordered intervals $I_0, \ldots, I_k$ is realized by k rating scales, so that the relaxed optimum Eq. (8) is attained by an actual design. Let the boundaries of the partition be $\tau_1 \leq \cdots \leq \tau_k$, so that $I_j = [\tau_j, \tau_{j+1})$ with $\tau_0 := 0$ and $\tau_{k+1} := 1$. For $i \in [k]$, let

$$\beta_i^*(\theta) := \mathbf{1}[\theta \geq \tau_i]$$

be the 0-1 step function with threshold τ_i. It is nondecreasing, hence an admissible scale. If $\theta \in I_j$, then $\tau_i \leq \theta$ exactly for the indices $i \leq j$. Hence exactly the scales $\beta_1^*, \ldots, \beta_j^*$ output a positive rating, and the item receives $n(\theta) = j$ positive ratings deterministically. The scales $(\beta_i^*)_{i \in [k]}$ therefore induce the assignment $\theta \mapsto j$ on I_j, and their objective equals $W_k(I_0, \ldots, I_k)$. Combined with Steps 1–3, this proves Theorem 4. For Kendall's τ function, the thresholds are $\tau_i = \frac{i}{k+1}$. □

A.5 Generalizing Theorem 4 to arbitrary score families

Step 2 in Appendix A.4 relied on the family $\{p(\theta)\}$ being stochastically nondecreasing, which is guaranteed by the monotonicity of the scales. In the adaptive setting of Section 6, the platform ranks items by an outcome in a finite set Ω according to an order $\prec$ of its choice. The resulting family of outcome distributions need not be monotone with respect to $\prec$. For instance, the number of positive responses along a path of the tree is not a monotone function of the quality in general. The following lemma shows that for weight functions such as Kendall's τ, the conclusion of Steps 1 and 2 holds for *every* family, monotone or not.

Lemma 11 (Relaxation to ordered intervals)**.** *Let S be a finite set with a total order $\prec$. For $s, s' \in S$, write $\operatorname{sign}_\prec(s, s') := 1$ if $s \prec s'$, $\operatorname{sign}_\prec(s, s') := -1$ if $s' \prec s$, and $\operatorname{sign}_\prec(s, s') := 0$ if $s = s'$. Let w be a weight function that is nonincreasing in θ_1 and nondecreasing in θ_2 on $\{\theta_1 < \theta_2\}$. Then for every measurable map $\theta \mapsto p(\theta) \in \Delta(S)$,*

$$2 \int_{\theta_1 < \theta_2} w(\theta_1, \theta_2) \sum_{s, s' \in S} \operatorname{sign}_\prec(s, s')\, p_s(\theta_1)\, p_{s'}(\theta_2)\, d\theta_1\, d\theta_2 \;\leq \sup_{I_0, \ldots, I_{|S|-1}} W(I_0, \ldots, I_{|S|-1}),$$

where the supremum is over partitions of $[0, 1]$ into $|S|$ ordered intervals and $W(I_0, \ldots, I_{|S|-1})$ is given by Eq. (9) with $k + 1$ replaced by $|S|$. In particular, the right side does not depend on the order $\prec$, and an optimal design never benefits from randomizing over orders.

Proof. Relabel the elements of S as $0 \prec 1 \prec \cdots \prec |S| - 1$, so that $\operatorname{sign}_\prec(s, s') = \operatorname{sign}(s' - s)$. Let $K := |S| - 1$. The left side is then Eq. (2) with the score set $\{0, 1, \ldots, k\}$ replaced by $\{0, 1, \ldots, K\}$, and we continue to write $W_k(p)$ for it. The proof replaces the quantile coupling of Step 2 by a discretization, a convexity argument and an exchange argument.

Reduction to finitely many qualities. Fix $N \geq 1$ and split $[0,1]$ into the cells $C_\alpha := [\frac{\alpha-1}{N}, \frac{\alpha}{N})$ for $\alpha \in [N]$. Let $p^{(N)}$ be the map that is constant on each cell and equal there to the average of p over the cell. By the Lebesgue differentiation theorem, $p^{(N)}(\theta) \to p(\theta)$ for almost every θ. All entries of p and $p^{(N)}$ lie in $[0,1]$ and w is integrable, so dominated convergence applied to Eq. (2) gives $W_k(p^{(N)}) \to W_k(p)$. Hence it suffices to prove the lemma for maps p that are constant on the N cells, for every N. Such a map is a point $(p^1, \dots, p^N)$ in the product of N simplices, where $p^\alpha \in \Delta(\{0, 1, \dots, K\})$ is the value of p on C_α.

Multilinearity and reduction to deterministic maps. For such a map, Eq. (2) splits into pairs of qualities in different cells and pairs in the same cell:

$$W_k(p) = 2 \sum_{\alpha<\beta} \Big(\int_{C_\alpha \times C_\beta} w \Big) \sum_{i,j} \operatorname{sign}(j-i)\, p_i^\alpha p_j^\beta + 2 \sum_{\alpha} \Big(\int_{(C_\alpha \times C_\alpha) \cap \{\theta_1 < \theta_2\}} w \Big) \sum_{i,j} \operatorname{sign}(j-i)\, p_i^\alpha p_j^\alpha.$$

The second sum vanishes, because by antisymmetry of $\operatorname{sign}(j-i)$ every distribution p satisfies

$$\sum_{i,j} \operatorname{sign}(j-i)\, p_i p_j = 0.$$

Hence $W_k(p)$ is a multilinear function of $(p^1, \dots, p^N)$: fixing all p^β with $\beta \neq \alpha$, it is an affine function of p^α. An affine function on a simplex attains its maximum at a vertex. Therefore, for $\alpha = 1, 2, \dots, N$ in turn, we may replace p^α by a point mass δ_{n_α} without decreasing W_k, where

$$n_\alpha \in \operatorname*{argmax}_i W_k(p^1, \dots, p^{\alpha-1}, \delta_i, p^{\alpha+1}, \dots, p^N).$$

After N replacements, the map is a deterministic sequence $(n_1, \dots, n_N) \in \{0, 1, \dots, K\}^N$ whose objective

$$V(n_1, \dots, n_N) = 2 \sum_{\alpha<\beta} \Big(\int_{C_\alpha \times C_\beta} w \Big) \operatorname{sign}(n_\beta - n_\alpha)$$

is at least $W_k(p)$. The same convexity argument shows that randomizing the order $\prec$ cannot help. If the platform ranked s above s' with some probability, the objective would be affine in that probability, so one of the two deterministic choices is at least as good.

Exchange argument. Call a pair of cells $\alpha < \beta$ with $n_\alpha > n_\beta$ an *inversion*. Among the finitely many sequences that maximize V over $\{0, 1, \dots, K\}^N$, choose one with the fewest inversions. We show that it has none. Suppose for contradiction that $\alpha < \beta$ is an inversion, with $a := n_\alpha > b := n_\beta$. Consider the sequence obtained by *swapping* the two values, $n'_\alpha := b$, $n'_\beta := a$, and $n'_\gamma := n_\gamma$ otherwise. We compare the contributions to V before and after the swap, pair by pair. Only pairs involving α or β change.

- *The pair* (α, β). Its contribution changes from

$$-2 \int_{C_\alpha \times C_\beta} w \quad \text{to} \quad +2 \int_{C_\alpha \times C_\beta} w,$$

  so it weakly increases.

- *A cell* γ *with* $\alpha < \gamma < \beta$*, with value* $c := n_\gamma$. The pairs (α, γ) and (γ, β) contribute

$$2 \int_{C_\alpha \times C_\gamma} w \cdot \operatorname{sign}(c-a) + 2 \int_{C_\gamma \times C_\beta} w \cdot \operatorname{sign}(b-c)$$

  before the swap and

$$2 \int_{C_\alpha \times C_\gamma} w \cdot \operatorname{sign}(c-b) + 2 \int_{C_\gamma \times C_\beta} w \cdot \operatorname{sign}(a-c)$$

  after. Since $b < a$, we have $\operatorname{sign}(c-b) \geq \operatorname{sign}(c-a)$ and $\operatorname{sign}(a-c) \geq \operatorname{sign}(b-c)$, so each term weakly increases. This case uses only $w \geq 0$.

- *A cell γ with $\gamma > \beta$, with value c.* The pairs (α, γ) and (β, γ) contribute

$$2\int_{C_\alpha \times C_\gamma} w \cdot \operatorname{sign}(c-a) + 2\int_{C_\beta \times C_\gamma} w \cdot \operatorname{sign}(c-b)$$

before the swap and

$$2\int_{C_\alpha \times C_\gamma} w \cdot \operatorname{sign}(c-b) + 2\int_{C_\beta \times C_\gamma} w \cdot \operatorname{sign}(c-a)$$

after, so the change equals

$$2\Big(\int_{C_\alpha \times C_\gamma} w - \int_{C_\beta \times C_\gamma} w\Big)\big(\operatorname{sign}(c-b) - \operatorname{sign}(c-a)\big) \;\geq\; 0.$$

Here $\operatorname{sign}(c-b) \geq \operatorname{sign}(c-a)$ because $b < a$. The first factor is nonnegative because w is nonincreasing in its first argument. Indeed, for $\theta_1 \in C_\alpha$, $\theta_2 \in C_\beta$ and $\theta \in C_\gamma$ we have $\theta_1 < \theta_2 < \theta$, and hence $w(\theta_1, \theta) \geq w(\theta_2, \theta)$. Integrating over $\theta \in C_\gamma$ and then over $\theta_1 \in C_\alpha$ and $\theta_2 \in C_\beta$, which have the same length, gives

$$\int_{C_\alpha \times C_\gamma} w \geq \int_{C_\beta \times C_\gamma} w.$$

- *A cell γ with $\gamma < \alpha$, with value c.* Symmetrically, the change equals

$$2\Big(\int_{C_\gamma \times C_\beta} w - \int_{C_\gamma \times C_\alpha} w\Big)\big(\operatorname{sign}(a-c) - \operatorname{sign}(b-c)\big) \;\geq\; 0,$$

using $\operatorname{sign}(a-c) \geq \operatorname{sign}(b-c)$ and that w is nondecreasing in its second argument.

Hence the swap does not decrease the objective, so the swapped sequence is also optimal. It has strictly fewer inversions, for three reasons. First, the pair (α, β) is no longer inverted. Second, a cell γ between α and β with $n_\gamma \in [b, a]$ loses every inversion it had with α or β and gains none, since after the swap it sees the value $b \leq n_\gamma$ on its left and $a \geq n_\gamma$ on its right. Third, a cell between them with $n_\gamma \notin [b, a]$, and every cell outside the range from α to β, keeps the same number of inversions with $\{\alpha, \beta\}$, because the values a and b merely exchange places on the same side of γ. This contradicts the choice of a maximizer with the fewest inversions.

Conclusion. We have shown that for every N and every cell-constant p there is a nondecreasing sequence $(n_1, \ldots, n_N)$ whose objective is at least $W_k(p)$. Such a sequence is a partition of $[0, 1]$ into $K + 1$ ordered intervals with boundaries at multiples of $1/N$. Letting $N \to \infty$ proves the inequality of the lemma. The right side is the maximum of Eq. (9) over partitions, which involves the order only through the number $|S|$ of intervals. This proves the last assertion. □

Remark 12 (On the monotonicity assumption in Lemma 11)**.** *The third and fourth cases of the exchange argument are the only places that use the assumption that w is nonincreasing in θ_1 and nondecreasing in θ_2. Without it, Lemma 11 is false. For example, let $|S| = 2$, partition $[0, 1]$ into four quarters $A < B < C < D$, and let $w = 1$ on $A \times B$ and on $C \times D$ and $w = 0$ elsewhere. Consider the non-monotone deterministic assignment with $n = 0$ on $A \cup C$ and $n = 1$ on $B \cup D$. It ranks every pair in $A \times B$ and in $C \times D$ correctly and achieves $2 \cdot \left(\frac{1}{16} + \frac{1}{16}\right) = \frac{1}{4}$. In contrast, a partition into two ordered intervals ranks at most one of the two groups of pairs correctly and achieves at most $\frac{1}{8}$. This does not contradict Theorem 4: the assignment above is not stochastically nondecreasing, so by Eq. (3) it cannot be induced by nondecreasing scales. Indeed, for $k = 1$, the objective of a nondecreasing scale β under this w is*

$$\frac{1}{8}\left[(\bar{\beta}_B - \bar{\beta}_A) + (\bar{\beta}_D - \bar{\beta}_C)\right] \leq \frac{1}{8},$$

where $\bar{\beta}_X$ denotes the average of β over X. The assumption of Lemma 11 holds for Kendall's τ function, for Spearman's ρ function, and more generally whenever w depends only on the gap $\theta_2 - \theta_1$ through a nondecreasing function, or only on θ_2 through a nondecreasing function. Whether Theorem 6 remains true for weight functions that violate it is an open question.

A.6 Proof of Theorem 6

We give the full proof of Theorem 6. Fix the number of ratings k and a weight function w that is nonincreasing in θ_1 and nondecreasing in θ_2, as in the statement of Theorem 6. Recall from Section 6 that an adaptive design consists of two components: a complete binary tree T of depth k whose internal nodes are labeled by nondecreasing scales, and a total order $\prec$ on the set of outcomes $\Omega = \{0,1\}^k$. Under T, an item of quality θ traverses the tree from the root. At a node labeled by scale β, it moves to the child labeled 1 with probability $\beta(\theta)$ and to the child labeled 0 otherwise, independently of all previous responses. The item's outcome $\omega(\theta) \in \Omega$ is the leaf it reaches, and we denote its distribution by $p(\theta) = (p_\omega(\theta))_{\omega\in\Omega} \in \Delta(\Omega)$. With $\mathrm{sign}_\prec$ as in Lemma 11, the platform objective is

$$W_k(T,\prec) = 2\int_{\theta_1<\theta_2} w(\theta_1,\theta_2) \sum_{\omega_1,\omega_2\in\Omega} \mathrm{sign}_\prec(\omega_1,\omega_2)\, p_{\omega_1}(\theta_1)\, p_{\omega_2}(\theta_2)\, d\theta_1\, d\theta_2. \tag{10}$$

This is Eq. (2) with the score set $\{0,1,\dots,k\}$, ordered by the usual order on integers, replaced by the outcome set Ω ordered by $\prec$. We follow the structure of the proof of Theorem 4 in Appendix A.4.

Why the quantile argument does not apply directly. The family $\{p(\theta)\}$ induced by a tree of nondecreasing scales is stochastically nondecreasing with respect to the *lexicographic* order on Ω. To see this, couple the responses of two items through common uniform random variables $U_1,\dots,U_k$ as in Appendix A.4. The item of higher quality receives a response at least as large at the root, and if the two responses are equal, both items proceed to the same child. By induction, the sequence of the higher-quality item is lexicographically at least as large. However, the platform may rank outcomes by an arbitrary order $\prec$, with respect to which the family need not be monotone. Step 2 of Appendix A.4 is therefore not available. We use Lemma 11 instead, which is why Theorem 6 is stated under the monotonicity assumption on w.

Steps 1 and 2: reduction to ordered intervals. Apply Lemma 11 with $S = \Omega$ and the order $\prec$. For every tree T,

$$W_k(T,\prec) \le \sup_{I_0,\dots,I_{2^k-1}} W(I_0,\dots,I_{2^k-1}),$$

where the supremum is over partitions of $[0,1]$ into 2^k ordered intervals and $W(I_0,\dots,I_{2^k-1})$ is given by Eq. (9) with 2^k intervals. By the last assertion of the lemma, the right side is the same for every order $\prec$, and randomizing over orders does not help either. Hence the maximum platform objective over all adaptive designs is at most the best objective of a partition into 2^k ordered intervals. In such a partition, all qualities in the ℓ-th interval are assigned a common outcome $\omega^{(\ell)}$, where $\omega^{(0)} \prec \omega^{(1)} \prec \cdots \prec \omega^{(2^k-1)}$.

Step 3: interval lengths. Step 3 of Appendix A.4 applies with 2^k intervals in place of $k+1$. For Kendall's τ function, the objective of ordered intervals of lengths $x_0,\dots,x_{2^k-1}$ is

$$1 - \sum_{\ell=0}^{2^k-1} x_\ell^2,$$

which is maximized when all $x_\ell = 2^{-k}$, giving the value $1-2^{-k}$. For a general weight function satisfying the assumption of the theorem, the interval boundaries minimize the within-interval weight in Eq. (9). The supremum is attained by the same compactness argument as in Appendix A.4.

Step 4: implementation by binary search. It remains to show that any partition of $[0,1]$ into 2^k ordered intervals $I_0,\dots,I_{2^k-1}$ is realized by a depth-k tree of nondecreasing rating scales together with a total order on Ω. Let the boundaries of the partition be $0 = t_0 \le t_1 \le \cdots \le t_{2^k} = 1$, so that $I_\ell = [t_\ell, t_{\ell+1})$. This is the only step that differs substantially from the non-adaptive setting, where a single layer of k step functions suffices.

Identify each outcome $\omega = (r_1,\dots,r_k) \in \{0,1\}^k$ with the integer

$$\sum_{j=1}^{k} r_j 2^{k-j} \in \{0,1,\dots,2^k-1\},$$

which is its value as a binary number whose most significant digit is r_1. Let $\prec$ be the induced order, which is the lexicographic order on response sequences. Consider a node at level $j \in [k]$ of the tree, reached by the responses $(r_1, \ldots, r_{j-1})$ of the first $j-1$ users, and let

$$m := \sum_{i=1}^{j-1} r_i 2^{k-i}.$$

We say that the intervals $I_m, I_{m+1}, \ldots, I_{m+2^{k-j+1}-1}$ are *consistent* with this node. They are the 2^{k-j+1} intervals whose index, written in binary with k digits, begins with $r_1, \ldots, r_{j-1}$. We label the node with the 0-1 step function

$$\beta_{(r_1,\ldots,r_{j-1})}(\theta) := \mathbf{1}\big[\theta \geq t_{m+2^{k-j}}\big],$$

whose threshold is the boundary between the lower half $I_m, \ldots, I_{m+2^{k-j}-1}$ and the upper half $I_{m+2^{k-j}}, \ldots, I_{m+2^{k-j+1}-1}$ of the consistent intervals. Each such scale is nondecreasing, hence admissible.

We claim that for every $j \in [k+1]$, an item with quality $\theta \in I_\ell$ reaches, after $j-1$ responses, the node whose consistent intervals contain I_ℓ. This holds for $j = 1$, since all intervals are consistent with the root. Suppose that it holds for j. At the node reached, the item is asked the question with threshold $t_{m+2^{k-j}}$. Since I_ℓ is one of the consistent intervals, $\theta \geq t_{m+2^{k-j}}$ if and only if I_ℓ lies in the upper half. Thus the response r_j is 1 exactly when the j-th binary digit of ℓ is 1, and the item proceeds to the child whose consistent intervals are the half containing I_ℓ. This proves the claim. For $j = k+1$, the only interval consistent with a leaf is the one indexed by its label, so the claim says that the item reaches the leaf $\omega = \ell$. Hence $\omega(\theta) = \ell$ deterministically for all $\theta \in I_\ell$. The tree therefore performs a binary search for the interval containing θ. It realizes the ordered-interval assignment from Step 2 with the outcomes in lexicographic order, so the relaxed optimum is attained by an admissible adaptive design.

For Kendall's τ function, the intervals have equal length 2^{-k} and $t_\ell = \ell 2^{-k}$. The binary search therefore uses threshold $\frac{1}{2}$ at the root and thresholds $\frac{1}{4}$ and $\frac{3}{4}$ at level 2. In general, the node reached by responses $(r_1, \ldots, r_{j-1})$ uses the threshold

$$t_{m+2^{k-j}} = \sum_{i=1}^{j-1} r_i 2^{-i} + 2^{-j}.$$

The outcome $\omega(\theta)$ then consists of the first k binary digits of θ, and $W_k^{\text{adaptive}} = 1 - 2^{-k}$. □

A.7 Partial adaptivity

Proposition 13 (Batch-adaptive scales)**.** *Let w be Kendall's τ function and let b divide k. Suppose the k ratings arrive in k/b batches of b ratings each. The b scales in a batch are fixed simultaneously but may depend on the responses in all previous batches. Items are ranked by the sequence $(c_1, \ldots, c_{k/b})$ of positive counts in each batch, according to a total order of the platform's choice. Then the maximum platform objective is $1 - \frac{1}{(b+1)^{k/b}}$.*

Proof. Each batch count satisfies $c_i \in \{0, 1, \ldots, b\}$, so the outcome set $S = \{0, 1, \ldots, b\}^{k/b}$ has $(b+1)^{k/b}$ elements. Kendall's τ function satisfies the assumption of Lemma 11. Hence, for every batch-adaptive design and every order on S, the objective is at most the best objective of a partition of $[0, 1]$ into $(b+1)^{k/b}$ ordered intervals. By Step 3 of Appendix A.4, this equals $1 - \frac{1}{(b+1)^{k/b}}$ and is attained only by intervals of equal length.

This value is achieved by a $(b+1)$-ary search, constructed as in Step 4 of the proof of Theorem 6 with the binary expansion replaced by the base-$(b+1)$ expansion. Index the equal-length intervals by $\ell \in \{0, 1, \ldots, (b+1)^{k/b} - 1\}$, and identify the outcome $(c_1, \ldots, c_{k/b})$ with the integer

$$\sum_{i=1}^{k/b} c_i (b+1)^{k/b-i},$$

ordered as integers. A node at round i reached by the counts $(c_1, \ldots, c_{i-1})$ is consistent with the $(b+1)^{k/b-i+1}$ intervals whose index in base $b+1$ begins with these digits. The b scales used in that

round are the 0-1 step functions whose thresholds split these consistent intervals into $b+1$ equal parts, exactly as in Step 4 of Appendix A.4 applied to the consistent block. An item in the c-th part then receives exactly c positive ratings in that round, so the batch count identifies the part containing θ. After k/b rounds, the item reaches the outcome indexed by its interval deterministically. For $b = k$, there is a single round and the construction reduces to Theorem 4, giving W_k^{multiple}. For $b = 1$, it is the binary search of Theorem 6, giving W_k^{adaptive}. □

A.8 Optimal threshold spacing for other weight functions

By Eqs. (8) and (9), together with Lemma 11 in the adaptive setting, the design problem reduces to choosing the boundaries $0 = t_0 \le t_1 \le \cdots \le t_n = 1$ of n ordered intervals. Here $n = k+1$ in the non-adaptive setting and $n = 2^k$ in the adaptive setting. The boundaries should minimize the total weight of the pairs of qualities that fall in the same interval. We give two examples.

- *Spearman's ρ function* $w(\theta_1, \theta_2) = \theta_2 - \theta_1$. The within-interval weight of an interval of length x is

$$\int_{0<u<v<x} (v-u)\, du\, dv = \frac{x^3}{6},$$

and the total weight is

$$\bar{w} = 2\int_{\theta_1<\theta_2} (\theta_2 - \theta_1)\, d\theta_1\, d\theta_2 = \frac{1}{3}.$$

Writing x_ℓ for the length of I_ℓ, Eq. (9) gives

$$W = \frac{1}{3} - \frac{1}{3}\sum_{\ell=0}^{n-1} x_\ell^3.$$

By convexity of $x \mapsto x^3$, this is maximized by equal lengths $x_\ell = 1/n$, with value $\frac{1}{3}\big(1 - \frac{1}{n^2}\big)$. Lagrange multipliers give the same conclusion, since they require all x_ℓ^2 to be equal. Hence under Spearman's ρ function the optimal thresholds are evenly spaced, exactly as under Kendall's τ function. The two objectives differ only in their value.

- *A weight function that only values the top half of the items*, $w(\theta_1, \theta_2) = \mathbf{1}[\theta_2 \ge \frac{1}{2}]$, which is constant in θ_1 and nondecreasing in θ_2. An interval contained in $[0, \frac{1}{2})$ has within-interval weight 0. An interval $[a, a+x]$ with $a \ge \frac{1}{2}$ has within-interval weight $x^2/2$. An interval $[a, t]$ with $a < \frac{1}{2} < t$ has within-interval weight

$$\int_{1/2}^{t} (\theta_2 - a)\, d\theta_2 = \tfrac{1}{2}\big(t - \tfrac{1}{2}\big)\big(t + \tfrac{1}{2} - 2a\big) \ge \tfrac{1}{2}\big(t - \tfrac{1}{2}\big)^2,$$

which is at least the weight of $[\frac{1}{2}, t]$, because the pairs (θ_1, θ_2) in the interval with $\theta_1 < \frac{1}{2} \le \theta_2$ tie. Hence it is never worse to place a boundary exactly at $\frac{1}{2}$, and intervals below $\frac{1}{2}$ beyond the first are wasted. The optimal partition therefore uses one interval $[0, \frac{1}{2})$ and splits $[\frac{1}{2}, 1]$ into $n-1$ intervals of equal length $\frac{1}{2(n-1)}$. The thresholds are $\tau_1 = \frac{1}{2}$ and $\tau_i = \frac{1}{2} + \frac{i-1}{2(n-1)}$ for $2 \le i \le n-1$, evenly spaced within the region that the platform cares about. Since the total weight is

$$\bar{w} = 2\int_{1/2}^{1} \theta_2\, d\theta_2 = \frac{3}{4},$$

the optimal value is $\frac{3}{4} - \frac{1}{4(n-1)}$. In particular, the thresholds are not evenly spaced in $[0, 1]$.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: All claims are theoretical and fully proven in the paper.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: All modeling assumptions that we make are in Section 3, and limitations are also discussed in Section 7.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: Proof outlines are in the main body and the full technical details are in the appendix.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [N/A]

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [N/A]

Justification: The paper does not include experiments.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [N/A]

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [N/A]

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [N/A]

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed the NeurIPS Code of Ethics and ensured the paper conforms to it.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper clearly states and theoretically proves the benefit of allowing heterogeneous rating scales for users, which is that items listed on online platforms can be distinguished more accurately. There is no negative societal impact to accurate ratings of platform listings.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper is theoretical and does not train models.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: The paper does not use existing assets.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLM is used only for editing and formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.