Your 2 is My 1, Your 3 is My 9: Handling Arbitrary Miscalibrations in Ratings

Jingyan Wang
Carnegie Mellon University
jingyanw@cs.cmu.edu

Nihar B. Shah
Carnegie Mellon University
nihars@cs.cmu.edu

ABSTRACT

A key step in building multi-agent systems is to gather data reported by the agents (people), in either cardinal (numeric ratings) or ordinal (rankings) form. Cardinal scores collected from people are well known to suffer from miscalibrations. A popular approach to address this issue is to assume simplistic models of miscalibration (such as linear biases) to de-bias the scores. This approach, however, often fares poorly because people's miscalibrations are typically far more complex and not well understood. *It is widely believed that in the absence of simplifying assumptions on the miscalibration, the only useful information in practice from the cardinal scores is the induced ranking.* In this paper we address the fundamental question of whether this widespread folklore belief is actually true. We consider cardinal scores with arbitrary (or even adversarially chosen) miscalibrations that is only required to be consistent with the induced ranking. We design rating-based estimators and prove that despite making no assumptions on the ratings, they strictly and uniformly outperform all possible estimators that rely on only the ranking. These estimators can be used as a plug-in to show the superiority of cardinal scores over ordinal rankings for a variety of applications, and we provide examples for A/B testing and ranking as a proof of concept. Our results thus provide novel fundamental insights in the eternal debate between cardinal and ordinal data.

KEYWORDS

Miscalibration; crowdsourcing; data collection methodologies; preference aggregation; multi-agent systems

ACM Reference Format:
Jingyan Wang and Nihar B. Shah. 2019. Your 2 is My 1, Your 3 is My 9: Handling Arbitrary Miscalibrations in Ratings. In *Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019,* IFAAMAS, 9 pages.

1 INTRODUCTION

> *"A raw rating of 7 out of 10 in the absence of any other information is potentially useless."* [24]
> *"The rating scale as well as the individual ratings are often arbitrary and may not be consistent from one user to another."* [1]

It is a common paradigm to evaluate and make decisions about a set of items, by soliciting and aggregating information from a heterogeneous group of people. One such example is conference peer review. Consider two items that need to be evaluated (for example, papers submitted to a conference) and two reviewers. Suppose each reviewer is assigned one distinct item for evaluation, and this assignment is done uniformly at random. The two reviewers provide their evaluations (say, in the range $[0, 1]$) for the respective item they evaluate, from which the better item must be chosen. However, the reviewers' rating scales may be miscalibrated. It might be the case that the first reviewer is lenient and always provides scores in $[0.6, 1]$ whereas the second reviewer is more stringent and provides scores in the range $[0, 0.4]$. Or it might be the case that one reviewer is moderate whereas the other is extreme – the first reviewer's 0.2 is equivalent to the second reviewer's 0.1 whereas the first reviewer's 0.3 is equivalent to the second reviewer's 0.9. More generally, the miscalibration of the reviewers may be arbitrary and unknown. Then is there any hope of identifying the better of the two items with any non-trivial degree of certainty?

A variety of applications involve collection and aggregation of human preferences or judgments in terms of cardinal scores (numeric ratings). A perennial problem with eliciting cardinal scores is that of miscalibration – the systematic errors introduced due to incomparability of cardinal scores provided by different people (see [18, 30] and references therein).

This issue of miscalibration is sometimes addressed by making simplifying assumptions about the form of miscalibration, and post-hoc corrections under these assumptions. Such models include one-parameter-per-reviewer additive biases [2, 16, 22, 28], two-parameters-per-reviewer scale-and-shift biases [28, 36] and others [14]. The calibration issues with human-provided scores are often significantly more complex causing significant violations to these simplified assumptions (see [18] and references therein). Moreover, the algorithms for post-hoc correction often try to estimate the individual parameters which may not be feasible due to low sample sizes. For instance, John Langford notes from his experience as the program chair of the ICML 2012 conference [21]:

> *"We experimented with reviewer normalization and generally found it significantly harmful."*

This problem of low sample size is exacerbated in a number of applications such as A/B testing where every reviewer evaluates only one item, thereby making the problem underdetermined even under highly restrictive models.

It is commonly believed that when unable or unwilling to make any simplifying assumptions on the bias in cardinal scores, the only useful information is the ranking of the scores [1, 15, 19, 24, 25, 35]. This perception gives rise to a second approach towards handling miscalibrations – that of using only the induced ranking or otherwise directly eliciting a ranking and not scores from the use. As noted by Freund et al. [15]:

> *"[Using rankings instead of ratings] becomes very important when we combine the rankings of many viewers who often*

Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), N. Agmon, M. E. Taylor, E. Elkind, M. Veloso (eds.), May 13–17, 2019, Montreal, Canada.

use completely different ranges of scores to express identical preferences."

These motivations have spurred a long line of literature on analyzing data that takes the form of partial or total rankings of items [1, 5, 9, 25, 33, 37, 40].

In this paper, we contest this widely held belief by addressing the following two **fundamental** questions:

- In the absence of simplifying modeling assumptions on the miscalibration, is there any estimator (based on the scores) that can outperform estimators based on the induced rankings?
- If only one evaluation per reviewer is available, and if each reviewer may have an arbitrary (possibly adversarially chosen) miscalibration, is there hope of estimation better than random guessing?

Our theory shows that the answer to both questions is "Yes". One need not make simplifying assumptions about the miscalibration and yet guarantee a performance superior to that of any estimator that uses only the induced rankings.

In more detail, we consider settings where a number of people provide cardinal scores for one or more from a collection of items. The calibration of each reviewer is represented by an unknown monotonic function that maps the space of true values to the scores given by this reviewer. These functions are arbitrary and may even be chosen adversarially. We present a class of estimators based on cardinal scores given by the reviewers which *uniformly* outperforms any estimator that uses only the induced rankings. A compelling feature of our estimators is that they can be used as a plug-in to improve ranking-based algorithms in a variety of applications, and we provide a theoretical proof-of-concept for two applications: A/B testing and ranking.

The techniques used in our analyses draw inspiration from the framework of Stein's shrinkage [20, 41] and empirical Bayes [34]. Our setting with 2 reviewers and 2 papers presented subsequently in the paper carries a close connection to the classic two-envelope problem (for a survey on the two-envelope problem, see [17]), and our estimator in this setting is similar in spirit to the randomized strategy [10] proposed by Thomas Cover. We discuss connections with the literature in more detail in Section 5.

Our work provides a new perspective on the eternal debate between cardinal scores and ordinal rankings. It is often believed that ordinal rankings are a panacea for the miscalibration issues with cardinal scores. Here we show that *ordinal estimators are not only statistically inadmissible (that is, Pareto-inefficient), they are also strictly and uniformly beaten by our cardinal estimators.* Our results thus uncover a new point on the tradeoff between cardinal and ordinal data collection. The fundamental theoretical results and insights established in this paper are envisaged to serve as a crucial building block towards the design of rating-based estimators under more benign assumptions on miscalibrations, and for more complex settings of data collection, in the future.

Finally, a note qualifying the scope of the problem setting considered here. In applications such as crowdsourced microtasks where workers often spend very little time answering every question, the cardinal scores elicited may not necessarily be consistent with the ordinal rankings, and moreover, ordinal rankings are often easier and faster to provide [37]. These differences cease to exist in a variety of applications such as peer review or in-person laboratory A/B tests which require the reviewers to spend a non-trivial amount of time and effort in the review process, and these applications form the motivation of this work.

An extended version of this paper is available on arXiv [44]. This extended version also contains complete proofs of all the results presented in the present paper, results for "ranking" metrics in addition to the metrics considered here, and additional simulations.

2 PRELIMINARIES

Consider a set of n items denoted as $\{1, \ldots, n\}$ or $[n]$ in short.[1] Each item $i \in [n]$ has an unknown value $x_i \in \mathbb{R}$. For ease of exposition, we assume that all items have distinct values. There are m reviewers $\{1, \ldots, m\}$ and each reviewer evaluates a subset of the items. The calibration of any reviewer $j \in [m]$ is given by an unknown, strictly-increasing function $f_j : \mathbb{R} \to \mathbb{R}$. (More generally, our results hold for any non-singleton intervals on the real line as the domain and range of the calibration functions). When reviewer j evaluates item i, the reported score is $f_j(x_i)$. We make no other assumptions on the calibration functions $f_1, \ldots, f_m$. We use the notation $>$ to represent a relative order of any items, for instance, we use "$1 > 2$" to say that item 1 has a greater value (ranked higher) than item 2. We assume that m and n are finite.

Every reviewer is assigned one or more items to evaluate. We denote the assignment of items to reviewers as $A = (S_1, \ldots, S_m)$, where $S_j \subseteq [n]$ is the set of items assigned to reviewer $j \in [m]$. We use the notation Π to represent the set of all permutations of n items. We let $\pi^* \in \Pi$ denote the ranking of the n items induced by their respective values $(x_1, \ldots, x_n)$, such that $x_{\pi^*(1)} > x_{\pi^*(2)} > \cdots > x_{\pi^*(n)}$. The goal is to estimate this underlying "true" ranking π^* from the evaluations of the reviewers. We consider two types of settings: an ordinal setting where estimation is performed using the rankings induced by each reviewer's reported scores, and a cardinal setting where the estimation is performed using the reviewers' scores (which can have an arbitrary miscalibration and only need to be consistent with the rankings). Formally:

- **Ordinal:** Each reviewer j reports a total ranking among the items in S_j, that is, the ranking of the items induced by the values $\{f_j(x_i)\}_{i \in S_j}$. An ordinal estimator observes the assignment A and the rankings reported by all reviewers.
- **Cardinal:** Each reviewer j reports the scores for the items in S_j, that is, the values of $\{f_j(x_i)\}_{i \in S_j}$. A cardinal estimator observes the assignment A and the scores reported by all reviewers.

The reader may observe that the setting described above considers "noiseless" data, where each reviewer reports either the scores $\{f_j(x_i)\}$ or the induced ranking. We also extend our results to a "noisy" setting later in the paper (Proposition 3.3) wherein the reported scores or rankings take the form $\{f_j(x_i) + \epsilon_{ij}\}$, where the noise terms $\{\epsilon_{ij}\}$ are i.i.d.

In order to compare the performance of different estimators, we use the notion of *strict uniform dominance*. Informally, we say that one estimator strictly uniformly dominates another if it incurs

[1] We use the standard notation of $[\kappa]$ to denote the set $\{1, \ldots, \kappa\}$ for any positive integer κ.

a strictly lower risk for all possible choices of the miscalibration functions and the item values.

In more detail, suppose that you wish to show that an estimator $\widehat{\pi}_1$ is superior to estimator $\widehat{\pi}_2$ with respect to some metric for estimating π^*. However, there is a clever adversary who intends to thwart your attempts. The adversary can choose the miscalibration functions of all reviewers and the values of all items, and moreover, can tailor these choices for different realizations of π^*. Formally, the adversary specifies a set of values $\{f_1^\pi, \ldots, f_m^\pi, x_1^\pi, \ldots, x_n^\pi\}_{\pi \in \Pi}$. The only constraints in this choice are that the miscalibration functions $f_1^\pi, \ldots, f_m^\pi$ must be strictly monotonic and that the item values $x_1^\pi, \ldots, x_n^\pi$ should induce the ranking π. In the sequel, we consider two ways of choosing the true ranking π^*: In one setting, π^* can be chosen by the adversary, and in the second setting π^* is drawn uniformly at random from Π. Once this ranking π^* is chosen, the actual values of the miscalibration functions and the item values are set as $f_1^{\pi^*}, \ldots, f_m^{\pi^*}$ and $x_1^{\pi^*}, \ldots, x_n^{\pi^*}$. The items are then assigned to reviewers according to the (possibly random) assignment A. The reviewers now provide their ordinal or cardinal evaluations as described earlier, and the two estimators $\widehat{\pi}_1$ and $\widehat{\pi}_2$ use these evaluations to compute their estimates. We say that estimator $\widehat{\pi}_1$ strictly uniformly dominates $\widehat{\pi}_2$, if $\widehat{\pi}_1$ is always guaranteed to incur a strictly smaller (expected) error than $\widehat{\pi}_2$. Formally:

Definition 2.1 (Strict uniform dominance). Let $\widehat{\pi}_1$ and $\widehat{\pi}_2$ be two estimators for the true ranking π^*. Estimator $\widehat{\pi}_1$ is said to strictly uniformly dominate estimator $\widehat{\pi}_2$ with respect to a given loss function $L : \Pi \times \Pi \to \mathbb{R}$ if

$$\mathbb{E}[L(\pi^*, \widehat{\pi}_1)] < \mathbb{E}[L(\pi^*, \widehat{\pi}_2)], \tag{1}$$

for all permissible $\{f_1^\pi, \ldots, f_m^\pi, x_1^\pi, \ldots, x_n^\pi\}_{\pi \in \Pi}$. The expectation is taken over any randomness in the assignment A and the estimators. If the true ranking π^* is drawn at random from a fixed distribution, then the expectation is also taken over this distribution; otherwise, inequality (1) must hold for all values of π^*.

Note that strict uniform dominance is a stronger notion than comparing estimators in terms of their minimax (worst-case) or average-case risks. Moreover, if an estimator $\widehat{\pi}_2$ is strictly uniformly dominated by some estimator $\widehat{\pi}_1$, then the estimator $\widehat{\pi}_2$ is statistically inadmissible (see [45, Definition 12.17] for a formal definition of statistical inadmissibility).

Finally, for ease of exposition, we focus on the 0-1 loss, $L(\pi^*, \pi) = \mathbb{1}\{\pi^* \neq \pi\}$. An extension to other metrics such as the Kendall-tau distance and the Spearman's footrule distance is provided in Appendix B of the extended version [44].

3 MAIN RESULTS

In this section we present our main theoretical results.

3.1 A canonical setting

We begin with a canonical setting that involves two items and two reviewers (that is, $n = 2$, $m = 2$), where each reviewer evaluates one of the two items. Our analysis for this setting conveys the key ideas underlying our general results. These ideas are directly applicable towards designing uniformly superior estimators for a variety of applications, and we subsequently demonstrate this general utility with two applications.

In this canonical setting, each of the two reviewers evaluates one of the two items chosen uniformly at random without replacement, that is, the assignment A is chosen uniformly at random from the two possibilities $(S_1 = 1, S_2 = 2)$ and $(S_1 = 2, S_2 = 1)$. Since each reviewer is assigned only one item, the ordinal data is vacuous. Then the natural ordinal baseline is an estimator which makes a guess uniformly at random:

$$\widehat{\pi}_{\text{can}}(A, \{\}) = \begin{cases} 1 \succ 2 & \text{with probability } 0.5 \\ 2 \succ 1 & \text{with probability } 0.5. \end{cases}$$

In the cardinal setting, let y_1 denote the score reported for item 1 by its respective reviewer, and let y_2 denote the score for item 2 reported by its respective reviewer. Since the calibration functions are arbitrary (and may be adversarial), it appears hopeless to obtain information about the relative ordering of x_1 and x_2 from just this data. Indeed, as we show below, standard estimators such as the sign test — ranking the items in terms of their reviewer-provided scores — provably fail to achieve this goal. More generally, the following theorem holds for all deterministic estimators, that is, estimators given by deterministic mappings from $\{A, y_1, y_2\}$ to the set $\{1 \succ 2, 2 \succ 1\}$.

Theorem 3.1. *No deterministic (cardinal or ordinal) estimator can strictly uniformly dominate the random-guessing estimator $\widehat{\pi}_{can}$.*

Proof. Let $\widehat{i}^{(1)} \in \text{argmax}_{i \in \{1,2\}} y_i$ denote the item which receives the higher score, and let $\widehat{i}^{(2)}$ denote the remaining item (with ties broken arbitrarily). First, we consider a deterministic estimator that always outputs $\widehat{i}^{(1)}$ as the item whose value is greater. We call this estimator the "sign estimator", denoted $\widehat{\pi}_{\text{sign}}$:

$$\widehat{\pi}_{\text{sign}}(A, y_1, y_2) = (\widehat{i}^{(1)} \succ \widehat{i}^{(2)}).$$

The proof consists of two steps. (1) We show that the sign estimator does not strictly uniformly dominate random guess. (2) Building on top of (1), we show that more generally, no deterministic estimator strictly uniformly dominates random guess.

Step 1: The sign estimator does not strictly uniformly dominate random guess.

We consider the following construction: let $x_1, x_2 \in (0, 1)$, and let $f_1(x) = x$ and $f_2(x) = x + 1$. Then the score given by reviewer 2 is higher than the score given by reviewer 1 regardless of the item values they are assigned. The sign estimator always observes $y_1 < y_2$, and outputs the item assigned to reviewer 2 as the greater item. Since the assignment is uniformly at random, the probability of error of the sign estimator is 0.5.

Step 2: No deterministic estimator strictly uniformly dominates random guess.

Let $\mathcal{A}$ be the set of the two assignments, $\mathcal{A} = \{(S_1 = 1, S_2 = 2), (S_1 = 2, S_2 = 1)\}$. A deterministic estimator $\widehat{\pi}_{\text{det}} : \mathcal{A} \times \mathbb{R} \times \mathbb{R} \to \{1 \succ 2, 1 \prec 2\}$ is a deterministic function that takes as input the assignment and the scores for the two items, and outputs the relative ordering of the two items. Step 1 has shown that the sign estimator does not strictly uniformly dominate random guess. Hence, we only need to prove that any deterministic estimator $\widehat{\pi}_{\text{det}}$ that is different from the sign estimator does not strictly uniformly dominate random guess. For this estimator $\widehat{\pi}_{\text{det}}$, there exist some input values $(a, \widetilde{y}_1, \widetilde{y}_2)$ such that the output of this deterministic estimator differs

from the sign estimator. If the two estimators $\widehat{\pi}_{\text{sign}}$ and $\widehat{\pi}_{\text{det}}$ only differ at points where $\widetilde{y}_1 = \widetilde{y}_2$, then we can use the same construction in Step 1 to show that the probability of error of $\widehat{\pi}_{\text{det}}$ is 0.5. Otherwise, there exist some input values where $\widetilde{y}_1 \neq \widetilde{y}_2$. Without loss of generality, assume $\widetilde{y}_1 > \widetilde{y}_2$. Then consider the following construction. Let $x_1 > x_2$. Let f_1, f_2 be strictly-increasing functions such that $f_1(x_1) = f_2(x_1) = \widetilde{y}_1, f_1(x_2) = f_2(x_2) = \widetilde{y}_2$. Regardless of the assignment, the score y_1 for item 1 is $\widetilde{y}_1$, and the score y_2 for item 2 is $\widetilde{y}_2$. Under assignment a, the deterministic estimator differs from the sign estimator, so the deterministic estimator gives the incorrect output ($1 \prec 2$). The assignment a happens with probability 0.5, so the probability of error of this deterministic estimator is at least 0.5. □

This theorem demonstrates the difficulty of this problem by ruling out all deterministic estimators. Our original question then still remains: is there any estimator that can strictly uniformly outperform the random-guessing ordinal baseline?

We show that the answer is yes, with the construction of a randomized estimator for this canonical setting, denoted as $\widetilde{\pi}^{\text{our}}_{\text{can}}$. This estimator is based on a function $w : [0, \infty) \to [0, 1)$ which may be chosen as any arbitrary strictly-increasing function. For instance, one could choose $w(x) = \frac{x}{1+x}$ or w as the sigmoid function. Given the scores y_1, y_2 reported for the two items, let $\widehat{i}^{(1)} \in \text{argmax}_{i \in \{1,2\}} y_i$ denote the item which receives the higher score, and let $\widehat{i}^{(2)}$ denote the remaining item (with ties broken uniformly). Then our randomized estimator outputs:

$$\widetilde{\pi}^{\text{our}}_{\text{can}}(A, y_1, y_2) = \begin{cases} \widehat{i}^{(1)} \succ \widehat{i}^{(2)} & \text{with probability } \frac{1+w(|y_1-y_2|)}{2} \\ \widehat{i}^{(2)} \succ \widehat{i}^{(1)} & \text{otherwise.} \end{cases} \tag{2}$$

Note that the the output of this estimator is independent of the assignment A, so in the remainder of this paper we also denote this estimator as $\widetilde{\pi}^{\text{our}}_{\text{can}}(y_1, y_2)$.

As an example, suppose that the values of the two items are $(x_1 = 4, x_2 = 7)$. Suppose the calibration function f_1 of reviewer 1 maps the values of these two items to $(f_1(x_1) = 1, f_1(x_2) = 5)$, and the calibration function f_2 of reviewer 2 maps them to $(f_2(x_1) = 6, f_2(x_2) = 8)$. Now, we observe the ratings $(y_1 = 1, y_2 = 8)$ with probability 0.5, in which case the estimator reports item 2 as greater with probability $\frac{1+w(7)}{2}$. With probability 0.5, we observe $(y_1 = 6, y_2 = 5)$, in which case the estimator reports item 2 as greater with probability $1 - \frac{1+w(1)}{2} = \frac{1-w(1)}{2}$. Since the function w is strictly-increasing, we have $w(7) > w(1)$. Using this fact and averaging the outcomes over these two cases yields a probability of success strictly greater than 0.5.

The following theorem now proves this result formally.

Theorem 3.2. *The randomized estimator $\widetilde{\pi}^{our}_{can}$ strictly uniformly dominates the random-guessing baseline $\widehat{\pi}_{can}$.*

Proof. We first re-write our estimator in (2) into an alternative and equivalent expression, and then prove the result on this new expression of our estimator.

We can split (2) into the following three cases, depending on the relative ordering of y_1 and y_2:

$$\widetilde{\pi}^{\text{our}}_{\text{can}}(A, y_1, y_2 \mid y_1 > y_2) = \begin{cases} 1 \succ 2 & \text{with probability } \frac{1+w(y_1-y_2)}{2} \\ 2 \succ 1 & \text{otherwise,} \end{cases} \tag{3a}$$

$$\widetilde{\pi}^{\text{our}}_{\text{can}}(A, y_1, y_2 \mid y_1 < y_2) = \begin{cases} 1 \succ 2 & \text{with probability } \frac{1-w(y_2-y_1)}{2} \\ 2 \succ 1 & \text{otherwise,} \end{cases} \tag{3b}$$

$$\widetilde{\pi}^{\text{our}}_{\text{can}}(A, y_1, y_2 \mid y_1 = y_2) = \begin{cases} 1 \succ 2 & \text{with probability } \frac{1}{2} \\ 2 \succ 1 & \text{otherwise.} \end{cases} \tag{3c}$$

Recall that the function w is from $[0, \infty)$ to $[0, 1)$. Now we define the following auxiliary function $\widetilde{w} : \mathbb{R} \to (0, 1)$:

$$\widetilde{w}(x) = \begin{cases} \frac{1+w(x)}{2} & \text{if } x > 0 \\ \frac{1}{2} & \text{if } x = 0 \\ \frac{1-w(-x)}{2} & \text{otherwise.} \end{cases} \tag{4}$$

Combining (3) and (4), we have

$$\widetilde{\pi}^{\text{our}}_{\text{can}}(A, y_1, y_2) = \begin{cases} 1 \succ 2 & \text{with probability } \widetilde{w}(y_1 - y_2) \\ 2 \succ 1 & \text{otherwise.} \end{cases} \tag{5}$$

Without loss of generality, assume $x_1 > x_2$. The assignment is either $a := (S_1 = 1, S_2 = 2)$ or $a' := (S_1 = 2, S_2 = 1)$ with probability 0.5 each. Thus, the estimator observes $(y_1 = f_1(x_1), y_2 = f_2(x_2))$ under assignment a, or $(y_1 = f_2(x_1), y_2 = f_1(x_2))$ under assignment a'. The probability of success of our estimator $\widetilde{\pi}^{\text{our}}_{\text{can}}$ is:

$$\begin{aligned} \mathbb{P}(\widetilde{\pi}^{\text{our}}_{\text{can}} = \pi^*) &= \sum_{\widetilde{a} \in \{a, a'\}} \mathbb{P}(\widetilde{\pi}^{\text{our}}_{\text{can}} = \pi^* \mid A = \widetilde{a}) \cdot \mathbb{P}(A = \widetilde{a}) \\ &\overset{\text{(i)}}{=} \frac{1}{2}\widetilde{w}(f_1(x_1) - f_2(x_2)) + \frac{1}{2}\widetilde{w}(f_2(x_1) - f_1(x_2)) \\ &\overset{\text{(ii)}}{=} \frac{1}{2}\left[1 + \widetilde{w}(f_1(x_1) - f_2(x_2)) - \widetilde{w}(f_1(x_2) - f_2(x_1))\right], \end{aligned} \tag{6}$$

where step (i) is true by plugging in (5), and step (ii) is true because $\widetilde{w}(x) + \widetilde{w}(-x) = 1$ by the definition of the function $\widetilde{w}$ in (4).

By the monotonicity of the functions f_1 and f_2, and by the assumption that $x_1 > x_2$, we have $f_1(x_1) + f_2(x_1) > f_1(x_2) + f_2(x_2)$, and therefore $f_1(x_1) - f_2(x_2) > f_1(x_2) - f_2(x_1)$. Since $w(0) \geq 0$ and w is monotonically increasing on $[0, \infty)$, it is straightforward to verify that $\widetilde{w}$ is monotonically increasing on $\mathbb{R}$. Hence, we have

$$\widetilde{w}(f_1(x_1) - f_2(x_2)) > \widetilde{w}(f_1(x_2) - f_2(x_1)) \tag{7}$$

Combining (6) and (7), we have

$$\mathbb{P}(\widetilde{\pi}^{\text{our}}_{\text{can}} = \pi^*) > 1/2.$$

□

The contrast between deterministic estimators and randomized estimators arises from the fact that a deterministic estimator "commits" to an action (deciding which item has a greater value). It performs well if the situation is aligned with this action (when the scores under miscalibration are consistent with the true ordering of the two items). However, due to its prior commitment it may fail if the situation is not aligned. In contrast, a randomized estimator balances out good and bad cases. The probability of the good case

(correct estimation) is greater than the probability of the bad case (incorrect estimation) for the randomized estimator (2), because it exploits the monotonic structure of the calibration functions, whereas this structure is lost in ordinal data.

While Theorem 3.2 considers a setting with "noiseless" observations (that is, where $y = f(x)$), an analogous result to Theorem 3.2 is established in the following proposition for a more general setting where the observations are noisy (with $y = f(x) + \text{noise}$). Formally, when reviewer $j \in [m]$ evaluates item $i \in [n]$, the reported score is $f_j(x_i) + \epsilon_{ij}$. We assume that the noise terms $\{\epsilon_{ij}\}_{i\in[n], j\in[m]}$ are drawn i.i.d. from any unknown distribution. In this setting of noisy reported scores, we modify Definition 2.1 of strict uniform dominance, and let the expectation include the randomness in the noise. The following theorem establishes the strict uniform dominance in the noisy setting for the cardinal estimator $\widetilde{\pi}^{\text{our}}_{\text{can}}$ in (2).

Proposition 3.3. *The canonical estimator $\widetilde{\pi}^{our}_{can}$ strictly uniformly dominates the random-guessing estimator $\widehat{\pi}_{can}$ in the presence of noise.*

Proof. The proof is a slight modification to the proof of Theorem 3.2. In Eq. (6) from the proof of Theorem 3.2, we replace all the noiseless terms $f_j(x_i)$ by the noisy terms $f_j(x_i) + \epsilon_{ij}$ for each $i \in \{1, 2\}$ and $j \in \{1, 2\}$. Taking an expectation over all the noise terms, we have

$$\begin{aligned}\mathbb{P}(\widetilde{\pi}^{\text{our}}_{\text{can}} = \pi^*) =& \frac{1}{2}\mathbb{E}_{\epsilon_{11},\epsilon_{12},\epsilon_{21},\epsilon_{22}}[\ 1 + \widetilde{w}((f_1(x_1) + \epsilon_{11}) - (f_2(x_2) + \epsilon_{22})) \\ & - \widetilde{w}((f_1(x_2) + \epsilon_{21}) - (f_2(x_1) + \epsilon_{12}))] \\ \overset{\text{(i)}}{=} & \frac{1}{2}\mathbb{E}_{\epsilon_1,\epsilon_2}[1 + \widetilde{w}(f_1(x_1) - f_2(x_2) + \epsilon_1 - \epsilon_2) \\ & -\widetilde{w}(f_1(x_2) - f_2(x_1) + \epsilon_1 - \epsilon_2)], \end{aligned} \tag{8}$$

where step (i) uses linearity of expectation with a change of variable names, as the noise terms ϵ_{ij} are i.i.d.

Without loss of generality, assume $x_1 > x_2$. Recall from the proof of Theorem 3.2 that $f_1(x_1) - f_2(x_2) > f_1(x_2) - f_2(x_1)$, and therefore we have the deterministic inequality

$$f_1(x_1) - f_2(x_2) + \epsilon_1 - \epsilon_2 > f_1(x_2) - f_2(x_1) + \epsilon_1 - \epsilon_2, \text{ for any } \epsilon_1, \epsilon_2 \in \mathbb{R}.$$

Using the monotonicity of $\widetilde{w}$, we have

$$\widetilde{w}(f_1(x_1) - f_2(x_2) + \epsilon_1 - \epsilon_2)) > \widetilde{w}(f_1(x_2) - f_2(x_1) + \epsilon_1 - \epsilon_2). \tag{9}$$

Taking an expectation over ϵ_1 and ϵ_2 in (9) and combining with (8) gives

$$\mathbb{P}(\widetilde{\pi}^{\text{our}}_{\text{can}} = \pi^*) > 1/2.$$

□

Observe that this result is quite general, since the noise distribution can be arbitrary and unknown.

3.2 A/B testing

We now demonstrate how to use the result in the canonical setting as a plug-in for more general scenarios. Specifically, we construct simple extensions to our canonical estimator, as a proof-of-concept for the superiority of cardinal data over ordinal data in A/B testing (this section) and ranking (Section 3.3). A/B testing is concerned with the problem of choosing the better of two given items, based on multiple evaluations of each item. In many applications of A/B testing, the two items are rated by disjoint sets of individuals (for example, when comparing two web designs, each user sees one and only one design). It is therefore important to take into account the different calibrations of different individuals, and this problem fits in our setting with $n = 2$ items and m reviewers. For simplicity, we assume that m is even. We consider an assignment obtained by assigning item 1 to some $m/2$ reviewers chosen uniformly at random (without replacement) from the set of m reviewers, and assigning item 2 to the remaining $m/2$ reviewers.[2]

For concreteness, we consider the following method of performing this random assignment. We first perform a uniformly random permutation of the m reviewers, and then assign the first $m/2$ reviewers in this permutation to item 1; we assign the last $m/2$ reviewers in this permutation to item 2. We let $y_1^{(1)}, \ldots, y_1^{(m/2)}$ denote the scores given by the $m/2$ reviewers to item 1, and let $y_2^{(1)}, \ldots, y_2^{(m/2)}$ denote the scores given by the $m/2$ reviewers assigned to item 2. Namely, the reviewers (in the permuted order) provide the scores $[y_1^{(1)}, \ldots, y_1^{(m/2)}, y_2^{(1)}, \ldots, y_2^{(m/2)}]$.

As in the canonical setting we studied earlier, in the absence of any direct comparison between the two items by the same reviewer, a natural ordinal estimator in the A/B testing setting is a random guess:

$$\widehat{\pi}_{\text{ab}}(A, \{\}) = \begin{cases} 1 \succ 2 & \text{with probability } 0.5 \\ 2 \succ 1 & \text{with probability } 0.5. \end{cases}$$

Now consider the following standard (deterministic) cardinal estimators:

- *Sign estimator:* The sign estimator outputs the item which has more pairwise wins:
  $\sum_{j=1}^{m/2} \mathbb{1}\{y_1^{(j)} > y_2^{(j)}\} \underset{2\succ1}{\overset{1\succ2}{\gtrless}} \sum_{j=1}^{m/2} \mathbb{1}\{y_2^{(j)} > y_1^{(j)}\}$.
- *Mean estimator:* The mean estimator outputs the item with the higher mean score:
  $\text{mean}(y_1^{(1)}, \ldots, y_1^{(m/2)}) \underset{2\succ1}{\overset{1\succ2}{\gtrless}} \text{mean}(y_2^{(1)}, \ldots, y_2^{(m/2)})$.
- *Median estimator:* The median estimator outputs the item with the higher median score: $\text{median}(y_1^{(1)}, \ldots, y_1^{(m/2)}) \underset{2\succ1}{\overset{1\succ2}{\gtrless}} \text{median}(y_2^{(1)}, \ldots, y_2^{(m/2)})$.

In each estimator, ties are assumed to be broken uniformly at random.

We now show that despite using the scores given by all m reviewers, where m can be arbitrarily large, these natural estimators fail to uniformly dominate the naïve random-guessing ordinal estimator $\widehat{\pi}_{\text{ab}}$.

Theorem 3.4. *For any (even) number of reviewers, none of the sign, mean, and median estimators can strictly uniformly dominate the random-guessing estimator $\widehat{\pi}_{ab}$.*

For Theorem 3.4 and all results to follow, we provide sketches of the proofs in the present paper, and refer the reader to Section 5 of the extended version [44] for the complete proofs.

[2] Our results also hold in the following settings: (a) Each reviewer is assigned one of the two items independently and uniformly at random. (b) Reviewers are grouped (in any arbitrary manner) into $m/2$ pairs, and within each pair, the two reviewers are assigned one distinct item each uniformly at random.

Proof sketch. We give a construction where the mean, median and sign estimators have a probability of error of 0.5. Let the item values be bounded as $x_1, x_2 \in (0,1)$, and let the m reviewer calibration functions be as follows:

$$f_j(x) = \begin{cases} x + (j-1) & \text{if } 1 \le j \le m-1 \\ x + \dfrac{m(m-1)}{2} & \text{if } j = m. \end{cases}$$

In this construction, one reviewer (specifically, reviewer m) has a significantly greater bias than the rest of the reviewers, and the ranges of the calibration functions are disjoint.

For the mean estimator, it can be verified that an item has a greater sum of scores if and only if reviewer m is assigned to this item. By symmetry of the assignment, the mean estimator makes an error with probability 0.5.

For the median estimator and the sign estimator, since the ranges of the calibration functions are disjoint, it can be verified that the output of the median and sign estimators only depend on the assignment, regardless of the two item values. Again, by symmetry of the assignment, it is equally likely for any set of $m/2$ reviewers to be assigned to item 1 or item 2. Hence, the median and sign estimators make an error with probability 0.5. □

The negative result of Theorem 3.4 demonstrates the challenges even when one is allowed to collect an arbitrarily large number of scores for each item. We now build on top of our canonical estimator $\widetilde{\pi}^{\text{our}}_{\text{can}}$ from Section 3.1, and present a simple randomized estimator $\widetilde{\pi}^{\text{our}}_{\text{ab}}$ as follows:

(1) For every $j \in [m/2]$, use the canonical estimator $\widetilde{\pi}^{\text{our}}_{\text{can}}$ on the j^{th} pair of scores $(y_1^{(j)}, y_2^{(j)})$ and obtain the estimate $r_j := \widetilde{\pi}^{\text{our}}_{\text{can}}(y_1^{(j)}, y_2^{(j)}) \in \{1 \succ 2, 2 \succ 1\}$.
(2) Output the majority vote among the estimates $\{r_j\}_{j \in [m/2]}$ with ties broken uniformly at random.

The following theorem now shows that the results for the canonical setting from Section 3.1 translate to this A/B testing application.

Theorem 3.5. *For any (even) number of reviewers, the estimator $\widetilde{\pi}^{our}_{ab}$ strictly uniformly dominates the random-guessing estimator $\widehat{\pi}_{ab}$.*

Proof sketch. Consider any arbitrary values of items $x_1, x_2 \in \mathbb{R}$. By symmetry of the assignment, we apply Theorem 3.2 on each pair of scores $(y_1^{(j)}, y_2^{(j)})$ for $j \in [m/2]$, and show that the canonical estimator gives the correct output with probability strictly greater than 0.5 on each pair.

Now we show that combining the $m/2$ pairs by majority voting yields a probability of success strictly greater than 0.5. For each $j \in [m/2]$, define $V_j \in \{0,1\}$ as the indicator variable of the correctness of our canonical estimator on the j^{th} pair of scores. We set $V_j = 1$ if the canonical estimator gives the correct output on the j^{th} pair, and 0 otherwise. Then V_j is a Bernoulli random variable with mean strictly greater than 0.5. Moreover, the variables $\{V_j\}_{j=1}^{m/2}$ are independent given the item values.

Let $V = \sum_{j=1}^{m/2} V_j$ be the number of pairs for which the canonical estimator $\widetilde{\pi}^{\text{our}}_{\text{can}}$ gives the correct output. Recall that the majority voting procedure breaks ties uniformly at random. The probability of success of our estimator is

$$\mathbb{P}[V > m/4] + \frac{1}{2}\mathbb{P}[V = m/4].$$

It can be verified that this probability is strictly greater than $1/2$. □

This result thus illustrates the use of our canonical estimator $\widetilde{\pi}^{\text{our}}_{\text{can}}$ as a plug-in for A/B testing, and can be extended to the noisy setting in a similar fashion.

So far we have considered settings where there are only two items and where each reviewer is assigned only one item, thereby making the ordinal data vacuous. In the next section, we turn to an application that does not have these restrictions.

3.3 Ranking

It is common in practice to estimate the partial or total ranking for a list of items by soliciting ordinal or cardinal responses from individuals. In conference reviews, each reviewer is asked to rank [12, 38, 39] or rate [16, 39] a small subset of the papers, and this information is subsequently used to estimate a partial or total ranking of the papers. Applications for aggregating rankings also arise in voting [31, 47], peer grading [29] and meta-search [13]. Formally, we let $n > 2$ denote the number of items and m denote the number of reviewers. For simplicity, we focus on a setting where each reviewer reports noiseless evaluation of some pair of items, and the goal is to estimate the total ranking of all items. We consider a random design setup where the pairs compared are randomly chosen and randomly assigned to reviewers. We assume $1 < m < \binom{n}{2}$ so that the problem does not degenerate. Each reviewer evaluates a pair of items, and these pairs are drawn uniformly without replacement from the $\binom{n}{2}$ possible pairs of items. We let $A = (S_1, \ldots, S_m)$ denote these m pairs of items assigned to the m respective reviewers, where $S_j \in [n] \times [n]$ denotes the pair of items assigned to reviewer $j \in [m]$. For each pair $S_j = (i, i')$, denote the cardinal evaluation as $y(S_j) = (f_j(x_i), f_j(x_{i'}))$, and the ordinal evaluation as the induced ranking $b(S_j) \in \{i \succ i', i' \succ i\}$. Denote the set of ordinal observations as $\mathcal{B} = \{b(S_j)\}_{j=1}^m$, and the set of cardinal observations as $\mathcal{Y} = \{y(S_j)\}_{j=1}^m$. The inputs to an ordinal estimator are the reviewer assignment A and the ordinal information $\mathcal{B}$. The inputs to a cardinal estimator are the reviewer assignment A and the set of cardinal observations $\mathcal{Y}$. Finally, let $\mathcal{G}(\mathcal{B})$ denote a directed acyclic graph (DAG) with nodes comprising the n items and with an edge from any node i to any other node i' if and only if $\{i \succ i'\} \in \mathcal{B}$. One can see that under the current setup of the problem, the graph $\mathcal{G}(\mathcal{B})$ captures all requisite information in the ordinal observations. A topological ordering on $\mathcal{G}$ is any total ranking of its vertices which does not violate any pairwise comparisons indicated by $\mathcal{B}$.

We now present our (randomized) cardinal estimator $\widetilde{\pi}^{\text{our}}_{\text{rank}}(A, \mathcal{Y})$ in Algorithm 1. In words, this algorithm starts from any topological ordering of the items as the initial estimate of the true ranking. Then the algorithm scans one-by-one over the pairs whose items are adjacent in the initial estimated ranking. If a pair can be flipped (that is, if the ranking after flipping this pair is also a topological ordering), we uniformly sample a pair of scores for these two items from the cardinal observations $\mathcal{Y}$, and use the randomized estimator $\widetilde{\pi}^{\text{our}}_{\text{can}}$ to determine the relative ordering of this pair. After $\widetilde{\pi}^{\text{our}}_{\text{can}}$ is called, the positions of the two items in this pair are finalized. We

Algorithm 1: Our cardinal ranking estimator $\widetilde{\pi}^{\text{our}}_{\text{rank}}(A, \mathcal{Y})$.

1 Deduce the ordinal observations $\mathcal{B}$ from the cardinal observations $\mathcal{Y}$.
2 Compute a topological ordering $\widehat{\pi}$ on the graph $\mathcal{G}(\mathcal{B})$, with ties broken in order of the indices of the items.
3 $t \leftarrow 1$.
4 **while** $t < n$ **do**
5 Let $\widehat{\pi}_{\text{flip}}$ be the ranking obtained by flipping the positions of the t^{th} and the $(t+1)^{th}$ items in $\widehat{\pi}$.
6 **if** $\widehat{\pi}_{\text{flip}}$ is a topological ordering on $\mathcal{G}(\mathcal{B})$, and both the t^{th} and $(t+1)^{th}$ items are evaluated by at least one reviewer each in $\mathcal{Y}$ **then**
7 From all of the scores of the t^{th} item in $\mathcal{Y}$, sample one uniformly at random and denote it as $y_{\widehat{\pi}(t)}$. Likewise denote $y_{\widehat{\pi}(t+1)}$ as a randomly chosen score of the $(t+1)^{th}$ item from $\mathcal{Y}$.
8 Consider the two reviewers reporting the scores $y_{\widehat{\pi}(t)}$ and $y_{\widehat{\pi}(t+1)}$. Remove from $\mathcal{Y}$ all scores provided by these two reviewers.
9 **if** $\widetilde{\pi}^{\text{our}}_{\text{can}}(y_{\widehat{\pi}(t)}, y_{\widehat{\pi}(t+1)})$ outputs $\widehat{\pi}(t+1) \succ \widehat{\pi}(t)$ **then**
10 $\widehat{\pi} \leftarrow \widehat{\pi}_{\text{flip}}$.
11 **end**
12 $t \leftarrow t + 2$.
13 **else**
14 $t \leftarrow t + 1$.
15 **end**
16 **end**
17 Output $\widetilde{\pi}^{\text{our}}_{\text{rank}}(A, \mathcal{Y}) = \widehat{\pi}$.

remove all scores of these two reviewers from future use, and jump to the next pair that does not contain these two items.

The following theorem presents the main result of this section.

Theorem 3.6. *Suppose that the true ranking π^* is drawn uniformly at random from the collection of all possible rankings, and consider any ordinal estimator $\widehat{\pi}_{rank}$ for π^*. Then the cardinal estimator $\widetilde{\pi}^{our}_{rank}$ strictly uniformly dominates the ordinal estimator $\widehat{\pi}_{rank}$.*

Proof sketch. Since the prior distribution of the true ranking π^* is uniform, we show that an ordinal estimator is optimal for the 0-1 loss, if and only if the (possibly randomized) output of this ordinal estimator belongs to the set of all topological orderings with probability 1.

Now consider our cardinal estimator $\widetilde{\pi}^{\text{our}}_{\text{rank}}$ from Algorithm 1. We call a pair of items "flippable", if Algorithm 1 uses the canonical estimator to decide the relative ordering of this pair (that is, the if-condition in Line 6 in Algorithm 1 is true). If there exist no flippable pairs, then Algorithm 1 makes no change to the initial topological ordering $\widehat{\pi}$. We show that in this case, the cardinal estimator is equivalent to an optimal ordinal estimator. Now consider the case when there exists at least one flippable pair. It can be verified that this case happens with non-zero probability. Since the reviewers are assigned to items uniformly at random, we can apply Theorem 3.2 to each flippable pair. The probability that the canonical estimator $\widetilde{\pi}^{\text{our}}_{\text{can}}$ outputs the correct relative ordering of each flippable pair is strictly greater than 0.5. Finally, it can be verified that an improvement on the probability of correctness on each flippable pair translates to an improvement on the probability of success of the ranking. □

We note that Algorithm 1 runs in polynomial time (in the number of items n) because the two major operations of this estimator – finding a topological ordering, and checking if a ranking is a topological ordering on the DAG – can be implemented in polynomial time [11]. Theorem 3.6 thus demonstrates again the power of the canonical estimator $\widetilde{\pi}^{\text{our}}_{\text{can}}$ as a plug-in component to illustrate the superiority of cardinal data vs. ordinal data in a variety of applications. Extensions of our result to the Kendall-tau distance and the Spearman's footrule distance are presented in Appendix B of the extended version [44].

4 TRADEOFF BETWEEN ESTIMATION UNDER PERFECT CALIBRATION VS. MISCALIBRATION

In this section, we present a preliminary experiment showing the tradeoff between estimation under perfect calibration (all reviewers reporting the true values of the items) and estimation under miscalibration. For simplicity, we consider the canonical setting from Section 3.1. We evaluate the performance of our estimator under two scenarios: (1) perfect calibration, where $f_j(x) = x$ for $j \in \{1, 2\}$; (2) miscalibration with one biased reviewer, where $f_1(x) = x$ and $f_2(x) = x + 1$. We consider the function w in our estimator as $w(x) = \frac{\gamma x}{1+\gamma x}$, where $\gamma \in \{2^k \mid -10 \le k \le 10, k \in \mathbb{Z}\}$. We sample x_1 and x_2 uniformly at random from the interval $[0, 1]$.

The relative improvement $\rho_{\widehat{\pi}}(\widetilde{\pi})$ of an estimator $\widetilde{\pi}$ as compared to a baseline estimator $\widehat{\pi}$ is defined as: $\rho_{\widehat{\pi}}(\widetilde{\pi}) = \frac{\mathbb{E}[L(\pi^*, \widehat{\pi})] - \mathbb{E}[L(\pi^*, \widetilde{\pi})]}{\mathbb{E}[L(\pi^*, \widehat{\pi})]} \times 100\%$. A positive value of the relative improvement $\rho_{\widehat{\pi}}(\widetilde{\pi})$ indicates the superiority of estimator $\widetilde{\pi}$ over estimator $\widehat{\pi}$. Figure 1 shows the relative improvement of our estimator over the random-guessing baseline under perfect calibration and under miscalibration. Let us focus on a few regimes in this plot. First, when γ is close to 0, we have $w(x)$ close to 0. The estimator is close to random-guessing, corresponding to the left end of the curve. At the other extreme, when γ goes to infinity, we have $w(x)$ close to 1. The estimator always outputs the item with the higher score, and hence gives perfect estimation under perfect calibration. Under miscalibration, the biased reviewer always gives the higher score, and the estimator always chooses the item assigned to this biased reviewer. The probability of success of this estimator is the same as random guess, corresponding to the right end of the curve. Past the maximum point of the function at $\gamma = 1$, the value of the curve starts decreasing, suggesting a tradeoff of estimation accuracy under perfect calibration and under miscalibration. It is clear that points to the left of the maximum point are suboptimal, since there exist points with the same accuracy under miscalibration but improved accuracy under perfect calibration.

We thus see that robustness under arbitrary miscalibration comes at a cost of lower accuracy under perfectly calibration. Establishing a formal understanding of this tradeoff and designing estimators that are provably optimal (in terms of this tradeoff) are important open problems.

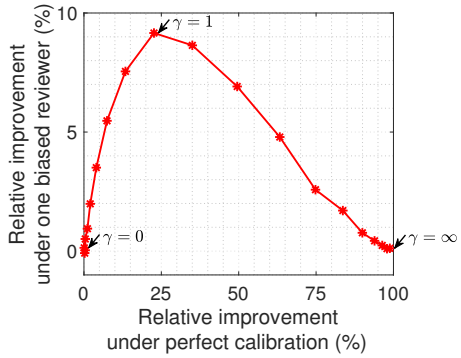

Figure 1: Relative improvement of our canonical estimator $\widetilde{\pi}^{\text{our}}_{\text{can}}$ under perfect calibration and under miscalibration of one biased reviewer, with $w(x) = \frac{\gamma x}{1+\gamma x}$. The error bars are too small to display.

5 CONNECTIONS TO THE LITERATURE

The canonical setting has a close connection to the randomized version of the two-envelope problem [10]. In the two-envelope problem, there are two arbitrary numbers. One of the two numbers is observed uniformly at random, and the other remains unknown. The goal is to estimate which number is larger. This problem can also be viewed from a game-theoretic perspective [17] as ours, where one player picks an estimator and the other player picks the two values. Cover [10] proposed a randomized estimator whose probability of success is strictly larger than 0.5 uniformly across all arbitrary pairs of numbers. The proposed estimator samples a new random variable Z whose distribution has a probability density function p with $p(z) > 0$ for all $z \in \mathbb{R}$. Then if the observed number is smaller than Z, the estimator decides that the observed number is the smaller number; if the observed number is larger than Z, the estimator decides that the observed number is the larger number.

Our canonical setting can be reduced to the two-envelope problem as follows. Consider the two values $f_1(x_1) - f_2(x_2)$ and $f_1(x_2) - f_2(x_1)$. Since the two items are assigned to the two reviewers uniformly at random, we observe one of these two values uniformly at random. By the assumption that f_1 and f_2 are monotonically increasing, we know that these two values are distinct, and furthermore, $f_1(x_1) - f_2(x_2) > f_1(x_2) - f_2(x_1)$ if and only if $x_1 > x_2$. Hence, the relative ordering of these two values is identical to the relative ordering of x_1 and x_2, reducing our canonical setting to the two-envelope problem. Our estimator $\widetilde{\pi}^{\text{our}}_{\text{can}}$ also carries a close connection to Cover's estimator to the two-envelope problem. Specifically, Cover's estimator can be equivalently viewed as being designated by a "switching function" [23]. This switching function specifies the probability to "switch" (that is, to guess that the unobserved value is larger), and is a monotonically decreasing function in terms of the observed value. The use of the monotonic function w in our estimator in (2) is similar in spirit.

Our original inspiration for our proposed estimator arose from Stein's phenomenon [41] and empirical Bayes [34]. This inspiration stems from the fact that the two items are not to be estimated in isolation, but in a joint manner. That said, a significant fraction of the work (e.g., [4, 8, 20, 34, 41, 43]) in these areas is based on deterministic estimators. In comparison, our negative result for all deterministic estimators (Theorem 3.1) and the positive result for our randomized estimator (Theorem 3.2) provide interesting insights in this space.

Broadly speaking, our work also shares similar motivation with incommensurable belief base merging in logic (e.g., [6, 7, 32]), and works in social choice theory that consider reviewer biases [27] or ordinal vs. cardinal data [3]. We take a statistical perspective and, motivated by challenges in peer review, focus on the setting where every reviewer only grades a small subset of the papers, and their grades share some extent of consistency (monotonicity).

6 CONCLUSIONS

Breaking the barrier of using only ranking data in the presence of arbitrary (and potentially adversarial) miscalibrations, we show that cardinal ratings can yield strict and uniform improvements over ordinal rankings. This result uncovers a novel, strictly-superior point on the tradeoff between cardinal scores and ordinal rankings, and provides a new perspective on this eternally debated tradeoff. Our (randomized) estimator allows for easily plugging into a variety of algorithms, thereby yielding it wide applicability.

In addition to the utility of cardinal ratings, the results of this paper provide an important **takeaway for practitioners**. In the application of conference peer review (which was a key motivation for this work), paper decisions are typically made in a deterministic fashion. However, our results suggest that for papers near the acceptance border, the difference in their scores is small, and could very well be due to issues of calibration of reviewers rather than inherent qualities of the papers. Our work thus suggests that a more fair alternative is to randomize the paper decisions at the border in a fashion along the lines of our proposed estimators in order to mitigate biases due to miscalibration.

This paper also leads to several **open problems**. First, while our estimators indeed uniformly outperform ordinal estimators, further improvements in our estimators (e.g., how to choose the function w in the canonical estimator, and how to design better estimators for A/B testing and ranking) may yield even better results. Second, it is of interest to obtain statistical bounds on the relative errors of the cardinal and ordinal estimators in terms of the unknown miscalibration functions. Third, although we consider the rating scales as continuous intervals, it is not hard to see that they extend to discrete scales as well (but with the strict inequality in Equation (1) sometimes replaced by a non-strict inequality to account for ties). Using our results to guide the choice of the scale used for elicitation is an open problem of interest. Finally, practical applications such as peer review do not suffer from the problem of miscalibration in isolation. It is a useful and challenging open problem to address miscalibration simultaneously with other issues such as noise [42], subjectivity [26] and strategic behavior [46].

ACKNOWLEDGMENTS

This work was supported in part by NSF grants CRII: CIF: 1755656 and CCF: 1763734. The authors thank Bryan Parno for very useful discussions on biases in conference peer review, and Pieter Abbeel for pointing out the related work on the two-envelope problem.

REFERENCES

[1] Ammar Ammar and Devavrat Shah. 2012. Efficient rank aggregation using partial data. In *SIGMETRICS/PERFORMANCE Joint International Conference on Measurement and Modeling of Computer Systems*. 355–366.

[2] Yukino Baba and Hisashi Kashima. 2013. Statistical Quality Estimation for General Crowdsourcing Tasks. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 554–562.

[3] Michel Balinski and Rida Laraki. 2007. A theory of measuring, electing, and ranking. *Proceedings of the National Academy of Sciences* 104, 21, 8720–8725.

[4] A. J. Baranchik. 1970. A Family of Minimax Estimators of the Mean of a Multivariate Normal Distribution. *Ann. Math. Statist.* 41, 2 (1970), 642–645.

[5] Jacob P. Baskin and Shriram Krishnamurthi. 2009. Preference aggregation in group recommender systems for committee decision-making. In *ACM Conference on Recommender Systems, RecSys*. 337–340.

[6] Salem Benferhat, Sylvain Lagrue, and Julien Rossit. 2007. An Egalitarist Fusion of Incommensurable Ranked Belief Bases under Constraints. In *AAAI Conference on Artificial Intelligence*. 367–372.

[7] Salem Benferhat, Sylvain Lagrue, and Julien Rossit. 2009. An Analysis of Sum-Based Incommensurable Belief Base Merging. In *International Conference on Scalable Uncertainty Management*. 55–67.

[8] M. E. Bock. 1975. Minimax Estimators of the Mean of a Multivariate Normal Distribution. *Ann. Statist.* 3, 1 (1975), 209–218.

[9] Wade D. Cook, Boaz Golany, Michal Penn, and Tal Raviv. 2007. Creating a consensus ranking of proposals from reviewers' partial ordinal rankings. *Computers & Operations Research* 34, 4 (2007), 954–965.

[10] Thomas M. Cover. 1987. *Pick the Largest Number*. Springer New York, New York, NY, 152–152.

[11] Sanjoy Dasgupta, Christos H. Papadimitriou, and Umesh Vazirani. 2008. *Algorithms* (1 ed.). McGraw-Hill, Inc.

[12] John R Douceur. 2009. Paper rating vs. paper ranking. *ACM SIGOPS Operating Systems Review* 43, 2 (2009), 117–121.

[13] Cynthia Dwork, Ravi Kumar, Moni Naor, and D. Sivakumar. 2001. Rank Aggregation Methods for the Web. In *International Conference on World Wide Web*. 613–622.

[14] Peter A. Flach, Sebastian Spiegler, Bruno Golénia, Simon Price, John Guiver, Ralf Herbrich, Thore Graepel, and Mohammed J. Zaki. 2010. Novel Tools to Streamline the Conference Review Process: Experiences from SIGKDD'09. *SIGKDD Explor. Newsl.* 11, 2 (2010), 63–67.

[15] Yoav Freund, Raj D. Iyer, Robert E. Schapire, and Yoram Singer. 2003. An Efficient Boosting Algorithm for Combining Preferences. *Journal of Machine Learning Research* 4 (2003), 933–969.

[16] Hong Ge, Max Welling, and Zoubin Ghahramani. 2013. A Bayesian model for calibrating conference review scores. (2013). http://mlg.eng.cam.ac.uk/hong/unpublished/nips-review-model.pdf [Online; accessed 03/01/2019].

[17] Alexander Gnedin. 2016. Guess the Larger Number. *preprint arXiv:1608.01899* (2016).

[18] Dale Griffin and Lyle Brenner. 2008. *Perspectives on Probability Judgment Calibration*. Wiley-Blackwell, Chapter 9, 177–199.

[19] Anne-Wil Harzing, Joyce Baldueza, Wilhelm Barner-Rasmussen, Cordula Barzantny, Anne Canabal, Anabella Davila, Alvaro Espejo, Rita Ferreira, Axele Giroud, Kathrin Koester, et al. 2009. Rating versus ranking: What is the best way to reduce response and language bias in cross-national research? *International Business Review* 18, 4 (2009), 417–432.

[20] William James and Charles Stein. 1961. Estimation with quadratic loss. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, Vol. 1. 361–379.

[21] John Langford. 2012. ICML acceptance statistics. (2012). http://hunch.net/?p=2517 [Online; accessed 05/14/2018].

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

[23] Mark D. McDonnell and Derek Abbott. 2009. Randomized switching in the two-envelope problem. In *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*. The Royal Society, 3309–3322.

[24] Ioannis Mitliagkas, Aditya Gopalan, Constantine Caramanis, and Sriram Vishwanath. 2011. User rankings from comparisons: Learning permutations in high dimensions. In *Allerton Conference on Communication, Control, and Computing*. 1143–1150.

[25] Sahand Negahban, Sewoong Oh, and Devavrat Shah. 2012. Iterative ranking from pair-wise comparisons. In *Advances in Neural Information Processing Systems*. 2474–2482.

[26] Ritesh Noothigattu, Nihar B. Shah, and Ariel Procaccia. 2018. Choosing how to choose papers. *arXiv preprint arxiv:1808.09057* (2018).

[27] António Osório. 2017. Judgement and ranking: living with hidden bias. *Annals of Operations Research* 253, 1 (2017), 501–518.

[28] S. R. Paul. 1981. Bayesian methods for calibration of examiners. *Brit. J. Math. Statist. Psych.* 34, 2 (1981), 213–223.

[29] Chris Piech, Jonathan Huang, Zhenghao Chen, Chuong Do, Andrew Ng, and Daphne Koller. 2013. Tuned models of peer assessment in MOOCs. *preprint arXiv:1307.2579* (2013).

[30] Robin S. Poston. 2008. Using and fixing biased rating schemes. *Commun. ACM* 51, 9 (2008), 105–109.

[31] Ariel D. Procaccia, Nisarg Shah, and Yair Zick. 2016. Voting rules as error-correcting codes. *Artif. Intell.* 231 (2016), 1–16.

[32] Guilin Qi, Weiru Liu, and David A. Bell. 2006. Merging Stratified Knowledge Bases under Constraints. In *AAAI Conference on Artificial Intelligence*. 281–286.

[33] Arun Rajkumar, Suprovat Ghoshal, Lek-Heng Lim, and Shivani Agarwal. 2015. Ranking from stochastic pairwise preferences: Recovering Condorcet winners and tournament solution sets at the top. In *International Conference on Machine Learning*. 665–673.

[34] Herbert Robbins. 1956. An empirical Bayes approach to statistics. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1. 157–163.

[35] Milton Rokeach. 1968. The Role of Values in Public Opinion Research. *Public Opinion Quarterly* 32, 4 (1968), 547–559.

[36] Magnus Roos, Jörg Rothe, and Björn Scheuermann. 2011. How to Calibrate the Scores of Biased Reviewers by Quadratic Programming. In *AAAI Conference on Artificial Intelligence*.

[37] Nihar B. Shah, Sivaraman Balakrishnan, Joseph Bradley, Abhay Parekh, Kannan Ramchandran, and Martin J. Wainwright. 2016. Estimation from pairwise comparisons: Sharp minimax bounds with topology dependence. *Journal of Machine Learning Research* 17, 1 (2016), 2049–2095.

[38] Nihar B. Shah, Joseph K Bradley, Abhay Parekh, Martin Wainwright, and Kannan Ramchandran. 2013. A case for ordinal peer-evaluation in MOOCs. In *NIPS Workshop on Data Driven Education*.

[39] Nihar B. Shah, Behzad Tabibian, Krikamol Muandet, Isabelle Guyon, and Ulrike Von Luxburg. 2017. Design and Analysis of the NIPS 2016 Review Process. *preprint arXiv:1708.09794* (2017).

[40] Nihar B. Shah and Martin J. Wainwright. 2018. Simple, Robust and Optimal Ranking from Pairwise Comparisons. *Journal of Machine Learning Research* (2018).

[41] Charles Stein. 1956. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1. 197–206.

[42] Ivan Stelmakh, Nihar B. Shah, and Aarti Singh. 2019. PeerReview4All: Fair and Accurate Reviewer Assignment in Peer Review. In *Algorithmic Learning Theory*.

[43] Kevin Tian, Weihao Kong, and Gregory Valiant. 2017. Learning Populations of Parameters. In *Advances in Neural Information Processing Systems*. 5778–5787.

[44] Jingyan Wang and Nihar B. Shah. 2018. Your 2 is My 1, Your 3 is My 9: Handling Arbitrary Miscalibrations in Ratings. *preprint arXiv:1806.05085* (2018).

[45] Larry Wasserman. 2010. *All of Statistics: A Concise Course in Statistical Inference*. Springer Publishing Company, Incorporated.

[46] Yichong Xu, Han Zhao, Xiaofei Shi, and Nihar B. Shah. 2018. On Strategyproof Conference Review. *arXiv preprint arxiv:1806.06266* (2018).

[47] H. P. Young. 1988. Condorcet's Theory of Voting. *American Political Science Review* 82, 4 (1988), 1231–1244.