Information Sciences 590 (2022) 322–345

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

An axiomatic distance methodology for aggregating multimodal evaluations

Adolfo R. Escobedo [a,*], Erick Moreno-Centeno [b], Romena Yasmin [a]

[a] School of Computing and Augmented Intelligence, Arizona State University, Tempe, AZ, USA
[b] Department of Industrial and Systems Engineering, Texas A&M University, College Station, TX, USA

ARTICLE INFO

Article history:
Received 17 May 2021
Received in revised form 28 December 2021
Accepted 30 December 2021
Available online 6 January 2022

Keywords:
Multimodal data aggregation
Group decision-making
Social choice
Axiomatic distances
Incomplete rankings and ratings

ABSTRACT

This work introduces a multimodal data aggregation methodology featuring optimization models and algorithms for jointly aggregating heterogeneous ordinal and cardinal evaluation inputs into a consensus evaluation. Specifically, this work derives mathematical modeling components to enforce three types of logical couplings between the collective ordinal and cardinal evaluations: Rating and ranking preferences, numerical and ordinal estimates, and rating and approval preferences. The proposed methodology is based on axiomatic distances rooted in social choice theory. Moreover, it adequately deals with highly incomplete evaluations, tied values, and other complicating aspects of group decision-making contexts. We illustrate the practicality of the proposed methodology in a case study involving an academic student paper competition. The methodology's advantages and computational aspects are further explored via synthetic instances sampled from distributions parametrized by ground truths and varying noise levels. These results show that multimodal aggregation effectively extracts a collective truth from noisy information sources and successfully captures the distinctive evaluation qualities of rating and ranking preference data.

1. Introduction

The natural tendency to reconcile multiple sources of conflicting information into a representative whole continues to spur the development of data aggregation techniques. Many such methods aim to systematically eliminate error/noise, thereby enhancing the quality of extractable information from individual information sources that collectively evaluate the same set of entities or systems of interest [1]. This motivation is especially prevalent in group decision-making, collective intelligence, and various other fields that seek to make sense of multiple subjective evaluations—judgments, preferences, estimates—which are inherently *heterogeneous* or contradictory. Furthermore, while most existing data aggregation methods work on a single data modality (continuous, ordinal, linguistic, etc.), there is a growing interest in developing new methodologies capable of integrating multiple data modalities (e.g. [2,3]). Integrating data with different modalities is important because different modalities are equipped to capture distinctive qualities of interest; hence, combining them could help extract more useful information than when separately considering each data modality.

In group decision-making and various other data aggregation applications, it is imperative that the aggregate evaluation is a good representation of the underlying "collective truth" inherent in the inputs. This overriding concern has led to

* Corresponding author at: PO Box 878809, Tempe, AZ 85287-8809, USA
E-mail addresses: adRes@asu.edu (A.R. Escobedo), emc@tamu.edu (E. Moreno-Centeno), ryasmin@asu.edu (R. Yasmin).

https://doi.org/10.1016/j.ins.2021.12.124
0020-0255/

increased attention on methods founded on the socio-theoretical concept of a *consensus* (e.g. [4–6]). Consensus aggregation methods find an aggregate evaluation of a set of *objects* that least disagrees (or, equivalently, most agrees) with the evaluations provided by a set of individuals or *judges*. Most methods differ by the evaluations' format (cardinal or ordinal) and the measure of disagreement used. In an ordinal evaluation, the objects are ordered based on their rank relative to a criterion of interest. For example, *ranking* vectors are used in group decision-making to sort the objects from "most preferred" to "least preferred" by assigning them a non-decreasing sequence of numerical values. In a cardinal evaluation, individual objects are given a *score* (a scalar value) quantifying the degree by which each possesses one or more qualities of interest (quantified according to an explicit or implicit reference scale). For example, *rating* vectors are used in group decision-making to record the scores of multiple objects; higher rating values (scores) typically indicate a higher quality/preference according to the evaluation criterion.

The usage of cardinal versus ordinal evaluations for making fair and effective collective decisions is a longstanding point of contention [7] dating back to the origins of voting theory [8]. On the one hand, advantages of ordinal evaluations include their avoidance of subjective scales, emphasis on pairwise comparisons, and their aggregation methods' robustness against outliers. On the other hand, advantages of cardinal evaluations include a lower cognitive load of elicitation (since reviewers can ostensibly evaluate each object independently), reflection of "intensities of preference" between objects, and their aggregation methods' computational efficiency. This work unifies these two contrasting theories by introducing a multimodal aggregation methodology founded on axiomatic distances. The proposed method is elaborated primarily in the context of group decision-making. However, the featured contributions have broader applicability because consensus aggregation methods are employed to reconcile heterogeneous evaluations in a variety of fields and applications, including bioinformatics [9], information retrieval [10], and wireless sensor networks [11], to name but a few.

This paper makes six main contributions:

- It introduces a general distance-based methodology for jointly aggregating ordinal and cardinal evaluations into a consensus multimodal evaluation.
- It derives mathematical modeling components to enforce three logical couplings between cardinal and ordinal evaluations.
- It constructs a ranking and rating consensus aggregation model that combines two axiomatic distances for group decision-making. Notably, the distances allow the inputs to be incomplete and contain ties, and they are generalizations of the Kemeny and Snell [12] ranking distance and the Cook and Kress [13] rating distance.
- It derives exact and approximate optimization models for solving the joint rating and ranking aggregation problem and computational enhancements based on polyhedral theory.
- It introduces supplementary techniques for identifying sources of high inconsistency in the evaluations and cases that may warrant further inspection.
- It assesses the practicality of the proposed methodology in a real-world case study and carefully-designed synthetic instances.

The paper is organized as follows: Section 2 provides an overview of axiomatic methodologies for ordinal and cardinal aggregation in group decision-making. This section also reviews approaches for incorporating subjective evaluations of multiple modalities. Section 3 describes the models and the axiomatic distances used to measure disagreement. Section 4 derives mathematical modeling components for enforcing three different logical interrelationships between ordinal and cardinal evaluations. Section 5 focuses on a multimodal aggregation problem associated with one of these couplings: the rating and ranking aggregation problem. Specifically, this section (i) proves the NP-hardness of the rating and ranking aggregation problem, (ii) derives a mixed-integer linear programming formulation and enhances it using polyhedral theory, (iii) derives a convexified formulation, and (iv) describes how to identify inconsistencies in the given evaluations. Section 6 illustrates the practicality of the proposed methodology in a case study involving the 2007 MSOM Student Paper Competition and in synthetic instances motivated by the case study and other practical considerations. Finally, Section 7 presents the conclusions of this work.

2. Literature Review

Group decision-making literature has addressed, for the most part, either the rankings-alone aggregation problem (e.g. [12,14–17]), or the ratings-alone aggregation problem (e.g. [18–23]). The ranking aggregation problem has been studied extensively, especially in the social choice literature. One of the most celebrated results is Arrow's impossibility theorem [14], which states that there is no "satisfactory" method to aggregate a set of rankings, where a satisfactory method is one that fulfills five conditions: universal domain, no imposition, monotonicity, independence of irrelevant alternatives, and non-dictatorship. In spite of this landmark result, different ranking aggregation methods have been developed to guarantee the fulfillment of a selected number of these and other desirable properties of the collective decision [6]. Kemeny and Snell [12] proposed a set of axioms that a distance metric between complete rankings should satisfy and proved that their distance uniquely satisfies all of them. More precisely, the distance is defined over weak orders, which are binary relations that are reflexive, transitive, and total. The distance measures the number of *rank reversals* between two rankings. A rank

reversal is incurred when two objects have a different relative order in the given rankings, and half of a rank reversal is incurred when two objects are tied in one ranking but not in the other. Kemeny and Snell defined the *median ranking* as the ranking that minimizes the sum of the distances to the input rankings. Their methodology has become synonymous with robust ranking aggregation. This axiomatic framework is known to ensure fairness, hinder manipulation, and mitigate individual bias in the aggregate outcome, as has been illustrated in different applications.

Bogart [24] extended the Kemeny and Snell distance framework to strict partial orders, which are binary relations that are irreflexive, asymmetric, and transitive. Cook et al. [25] developed a similar axiomatic distance measure for non-strict partial orders, which are binary relations that are reflexive, antisymmetric, and transitive; their measure is equivalent to the Kemeny and Snell and Bogart distances under the respective ranking subspaces. Subsequently, Hassanzadeh and Milenkovic [5] proposed a family of distance measures that prioritize the top part of the ranking and similarities between objects; these distances are founded on axioms similar to those of Kemeny and Snell. More recently, Moreno-Centeno and Escobedo [26] devised a generalization of the Kemeny-Snell ranking distance for incomplete rankings, which reduces to the original distance when the rankings are complete. In Yoo et al. [27], the authors bolstered the intuitiveness of this distance function by showing its connection to a generalization of the Kendall-τ ranking-correlation coefficient [28]. Unfortunately, the optimization problem that needs to be solved to find a median ranking via any of the measures mentioned above is NP-hard [15]. For this reason, a variety of algorithms have been developed to expedite the solution of the ranking aggregation problem (see [4]).

The difficulties presented by Arrow's impossibility theorem and the NP-hardness of finding a consensus ranking can be overcome by replacing ordinal rankings with (cardinal) ratings. Following this direction, Keeney [21] proved that the *averaging method* satisfies all of Arrow's desirable properties; in this method, the collective rating of an object is the average of the scores it receives. However, an immediate drawback of this approach is that it implicitly requires that all judges use the same rating scale; that is, all individuals must be equally strict or lenient. Such a standard is nearly impossible to enforce, even when providing detailed evaluation rubrics, as evidenced by the case study in Section 6. Moreover, the rating aggregation approach ignores the aspect of relative pairwise comparisons, which are fundamental towards avoiding specific undesirable outcomes—e.g., an object that would win a two-candidate election against every other object may not be selected as the winner [6].

The separation-deviation model of Hochbaum and Levin [20] overcomes the computational difficulties of the Kemeny-Snell model and mitigates the inadequacies of incomparable subjective scales. The model takes point-wise scores and potentially also pairwise comparison intensities as inputs, and it is one of the building blocks of the models proposed herein. When the chosen penalty functions are convex, the separation-deviation optimization problem is solvable in polynomial time.

Axiomatic distance methodologies have been extended to linguistic preferences [29], where instead of using numeric values (e.g., ratings or rankings), individual preferences are conveyed using a small predefined set of linguistic terms—a representative such set is {"very bad", "bad", "medium", "good", "very good"}. Very recently, Li et al. [30] proposed the first axiomatic distance on linguistic preferences for measuring group consensus. However, their consensus concept noticeably differs from the featured axiomatic distance approach. Expressly, Li et al. refer to a feedback adjustment mechanism through which judges iteratively discuss and modify their preferences based on the collective preference values until they reach a specific level of agreement. Moreover, although judges use members of the predefined linguistic set to express their subjective evaluations, these linguistic inputs are then transformed into scores based on a personalized numerical scale for each judge during the consensus reaching process. Accordingly, the collective preferences obtained suffer from similar drawbacks as the aforementioned cardinal aggregation methodologies.

In an attempt to circumvent the limitations associated with using a single modality of preferences, some recent works have proposed analyzing multiple modalities of subjective evaluations to arrive at a more comprehensive decision. There are two general approaches for utilizing such multimodal inputs: modality transformation and multi-objective optimization. An overview of these methods in the context of group decision-making can be found in Chen et al. [31].

The first primary approach involves transforming evaluations of various modalities into a single evaluation modality. As a representative example, Chiclana et al. [32] presented an aggregation process that allows inputs to be expressed in one of three modalities: preference orderings (i.e., rankings), utility functions (i.e., ratings), and fuzzy preference relations (binary relations that reflect the degree of preference of one object over another on a scale of 0 to 1). All three types of evaluations are converted into fuzzy preference relations. Finally, the concept of a fuzzy majority is used to aggregate the results and induce a global ranking of the alternatives by sorting the resulting score values obtained. A related but more inclusive approach in terms of allowable input modalities was introduced by Wu and Liao [33] which, in addition to the three modalities mentioned above, also allows interval-valued and linguistic preferences. This method consists of three steps. First, mutually exclusive groups of judges are formed based on preference modality utilized; second, all evaluations from a particular cluster are aggregated using methods appropriate to the associated modality; third, a ranking is induced from each cluster, and the resulting heterogeneous rankings are aggregated using a feedback mechanism similar to [30]. A significant drawback of works that align with this first primary approach is that modality transformations may lead to a loss of information [34].

The second primary approach for utilizing multimodal evaluations in group decision-making is integrating them through multi-objective optimization techniques. For instance, Fan et al. [35] proposed to aggregate multiplicative and fuzzy preference relations by minimizing the deviation between a solution priority vector (reflecting each object's priority weight collectively) and the input evaluations associated with each modality. The authors introduced a goal programming model that

reduces the two respective deviation functions into soft constraints. Wang et al. [36] introduced a chi-squared model that compares the squared deviation between a solution priority vector and multiplicative and fuzzy preference relations. The authors derived a single-objective nonlinear optimization model in which the two deviation functions are summed according to the weight assigned to the deviation from each judge. One drawback of these approaches is that, while the inputs can consist of multiple modalities, a single solution modality is explicitly optimized; this single modality may also be different from the input modalities. It is important to add that, although some of these models may allow judges to submit an evaluation consisting of multiple modalities, they do not explicitly consider or take advantage of this possibility.

Methods for utilizing multiple modalities of subjective evaluations have also been defined outside of the axiomatic group decision-making context. In crowdsourcing, Li et al. [3] proposed a statistical approach to jointly aggregate rating and ranking information gathered in two steps. In the first step, participants are asked to provide rating values for a large set of objects; these evaluations are aggregated into coarse aggregate scores. In the second step, participants are asked to rank a smaller set of alternatives that achieved the highest such scores. In the context of machine learning, Sader et al. [37] proposed integrating categorical data collected from experts and ranking data collected from novices to solve the ordinal classification problem. To mitigate computational difficulties encountered therein, Tang et al. [38,39] instead proposed a k-nearest neighbor-based classification approach. Although these machine learning models utilize multiple modalities, the nature of the underlying problems is inherently different from the proposed approach. The classification problem seeks to use the multimodal inputs to train a prediction model capable of correctly classifying additional evaluations that are not part of the training set. Furthermore, the axiomatic method presented herein seeks to derive a collective multimodal evaluation that provides a robust representation only of the given inputs; if more evaluations are received (even if they are duplicates of those already included), it is expected that the collective evaluation could be different. That said, our proposed approach could be adapted for use in ordinal classification. This is left for future research.

This paper addresses three main research gaps that can be observed in the preceding discussion:

- Existing multimodal aggregation methods explicitly or implicitly transform one or more modalities into a modality of a different type. Unfortunately, such transformations tend to lead to a loss of information [34]. Conversely, the proposed approach determines a consensus evaluation (which is multimodal and matches the input modalities) and utilizes distance functions specific to each input modality. Consequently, the solution does not need to transform one modality into another; instead, it interrelates the modalities through logical axioms.
- The group decision-making literature tends to assume that each judge expresses their evaluation in exactly one preference modality that aligns with the aggregation model. Accordingly, they do not tap into the rich information that could be derived from assessments in multiple modalities (e.g., a judge's rating evaluation whose values contradict the same judge's ranking evaluation). Furthermore, most works assume that the judges evaluate all objects, which is impractical in some real-world situations. The proposed methodology allows the evaluations to be expressed in cardinal and/or ordinal modalities, to be incomplete, and to contain ties.
- Alternative models that enable determining group consensus based on multimodal data stop at the aggregation process and do not consider additional uses for the multimodal data. This paper develops techniques for systematically identifying problematic evaluations by leveraging modality-specific inconsistencies with respect to the aggregate evaluation. For example, a judge's evaluation whose values highly contradict the collective assessment in one or two modalities could be grounds for initiating an investigation and/or further deliberations.

3. Preliminaries

This section introduces basic notation and definitions used throughout the paper. It reviews the concepts of the separation-deviation (SD) model, the axiomatic distance between incomplete ratings, and the axiomatic distance between incomplete rankings.

3.1. Basic Notation and Definitions

Let V be the universal set of n objects to be evaluated; without loss of generality, assign a unique identifier to each element so that $V = \{1, 2, \ldots, n\}$. There are m judges indexed by $k \in \{1, 2, \ldots, m\}$, each of who provides a (possibly incomplete) vector of scores or ratings, $\boldsymbol{a}^k$, over V. Specifically, a_j^k is the score given by judge k to object j, and a_j^k is undefined or assigned the token "•" if judge k did not rate object j; the subset of the objects rated by judge k is written as $V_{\boldsymbol{a}}^k \subseteq V$. It is also assumed that the input ratings may contain ties. Without loss of generality, the scores are contained in a prespecified interval $[L, U]$; the ratings' *range* is given by $R := U - L$. The *implied (cardinal) separation gap* (or *intensity of preference*) of object i to object j expressed by judge k is

$$p_{ij}^k = \begin{cases} a_i^k - a_j^k & \text{if } \; i \in V_{\boldsymbol{a}}^k \text{ and } j \in V_{\boldsymbol{a}}^k \\ \text{undefined} & \text{otherwise.} \end{cases} \tag{1}$$

Judge k may also provide a (possibly incomplete) ranking vector, $\boldsymbol{b}^k$, over V. Specifically, b_j^k is the rank position (an ordinal number) given by judge k to object j, and b_j^k is undefined or assigned the token "$\bullet$" if judge k did not rank object j; the subset of the objects ranked by judge k is written as $V_{\boldsymbol{b}}^k \subseteq V$. This work also assumes that the input rankings may contain ties according to the following convention. When $\boldsymbol{b}$ ties all the objects in a subset $V'_{\boldsymbol{b}} \subseteq V_{\boldsymbol{b}}$ and these objects are all ranked strictly worse than $(p-1)$ other objects in $V_{\boldsymbol{b}}$, where $p \geq 1$, then $b_i = p$ for all $i \in V'_{\boldsymbol{b}}$. Likewise, an object $j \in V_{\boldsymbol{b}} \setminus V'_{\boldsymbol{b}}$ that holds the next (worse) ranking position relative to $i \in V'_{\boldsymbol{b}}$ receives the rank $b_j = p + |V'_{\boldsymbol{b}}|$. Stated otherwise, the expression $(|V_{\boldsymbol{b}}^k| - b_i^k)$ reflects how many objects from $V_{\boldsymbol{b}}^k$ are tied or ranked worse than i (excluding itself).

The *implied (ordinal) separation gap* (or *preference*) of object i to object j expressed by judge k is given by $\text{sign}(b_i^k - b_j^k)$, if judge k ranks both objects, and is undefined otherwise. The sign function is defined as

$$\text{sign}(x) = \begin{cases} -1 & \text{if } x < 0, \\ 0 & \text{if } x = 0, \\ 1 & \text{if } x > 0. \end{cases} \tag{2}$$

Although we assume that every judge gives both a rating and a ranking, the herein proposed methodology also applies to situations where not all judges provide both types of evaluations. Consequently, our methodology is also applicable to cases where some judges give only ratings and other judges give only rankings.

Since the axiomatic distances used in the proposed methodology rely on comparing the scores or ranks of pairs of objects, it is convenient to define the *pairwise comparison arc set* over the object set $V_{\boldsymbol{a}}^k$ (or $V_{\boldsymbol{b}}^k$) as follows:

$$\mathcal{A}^k := \{(i,j) : i \in V_{\boldsymbol{a}}^k, j \in V_{\boldsymbol{a}}^k, i < j\}, \tag{3}$$

($\mathcal{B}^k$ is defined analogously for the ranking data); by convention, the pairwise comparison arc sets contain only arcs from lower to higher indices to eliminate duplicate comparison pairs.

Given a rating vector $\boldsymbol{a}$, we denote as $\text{rank}(\boldsymbol{a})$ the ranking obtained by first sorting the objects by their non-decreasing scores in $\boldsymbol{a}$ and then assigning to each object the ordinal number corresponding to its position in the sorted list. For example, $\boldsymbol{a} = (4.5, 5, 2, \bullet, 2, 1.7)$ induces $\text{rank}(\boldsymbol{a}) = (2, 1, 3, \bullet, 3, 5)$. Throughout the paper, it is assumed that higher ratings are superior (preferred) to lower ratings—as is customary in product reviews, student grading, and other group decision-making contexts.

A distance-based *consensus* is defined as the optimal solution to the cardinal aggregation (CA) problem or the ordinal aggregation (OA) problem, which can be respectively written succinctly as

$$\text{(CA)} \quad \min_{\boldsymbol{x}} \sum_{k=1}^{m} d(\boldsymbol{x}, \boldsymbol{a}^k) \quad || \quad \text{(OA)} \quad \min_{\boldsymbol{y}} \sum_{k=1}^{m} d(\boldsymbol{y}, \boldsymbol{b}^k), \tag{4}$$

where the respective solution vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ are assumed to be complete. This work also assumes that the solution vectors may contain ties.

3.2. Review of the Separation-Deviation Model

The separation-deviation (SD) model can be applied to group decision-making problems where the input is given as pairwise comparisons and/or point-wise scores. In the model, the variable x_i is the i^{th} object's aggregate score and, thus, $(x_i - x_j)$ represents the aggregate separation gap of the i^{th} over the j^{th} object. A set of separation gaps p_{ij} given as inputs must be *consistent*, that is, for all triplets (h,i,j), $p_{hi} + p_{ij} = p_{hj}$. The consistency of a set of separation gaps is equivalent to the existence of a set of scores ω_i for $i = 1, \ldots, n$ such that $p_{ij} = \omega_i - \omega_j$ [20]. The mathematical programming formulation for the SD problem is as follows:

$$(SD) \qquad \min_{\boldsymbol{x}} \sum_{k=1}^{m} \sum_{i=1}^{n} \sum_{j=1}^{n} f_{ij}^k\Big((x_i - x_j) - p_{ij}^k\Big) + \sum_{k=1}^{m} \sum_{i=1}^{n} g_i^k(x_i - a_i^k), \tag{5a}$$

$$\text{subject to} \quad L \leqslant x_i \leqslant U \qquad i = 1, \ldots, n, \tag{5b}$$

$$x_i \in \mathbb{Z} \qquad i = 1, \ldots, n. \tag{5c}$$

The function $f_{ij}^k(\cdot)$ penalizes the difference between the aggregate separation gap and the k^{th} reviewer's separation gap for object-pair (i,j). The function $g_i^k(\cdot)$ penalizes the difference between the aggregate score of i and the k^{th} reviewer's score of i. In order to ensure polynomial-time solvability, $f_{ij}^k(\cdot)$ and $g_i^k(\cdot)$ must be convex. In the context of rating aggregation, the penalty functions assume the value 0 for the argument 0; this means that if the output separation gap for object-pair (i,j) (given by $(x_i - x_j)$) agrees with p_{ij}^k, then $f_{ij}^k((x_i - x_j) - p_{ij}^k) = f_{ij}^k(0) = 0$. If $i \notin V_{\boldsymbol{a}}^k$, then $g_i^k(\cdot)$ is set to the constant function 0; similarly, if

$i \notin V_{\boldsymbol{a}}^k$ or $j \notin V_{\boldsymbol{a}}^k$, then $f_{ij}^k(\cdot)$ is set to the constant function 0. Furthermore, for linear $f_{ij}^k(\cdot)$ and $g_i^k(\cdot)$ with $L, U \in \mathbb{Z}$, the resulting problem can be solved as a linear program and $\boldsymbol{x}$ is guaranteed to be integral due to the unimodularity of the constraint coefficient matrix.

The SD problem is a special case of the convex dual of the minimum cost network flow (CDMCNF) problem [20]. The most efficient algorithm known for the CDMCNF has a running time of $O(mn \log \frac{n^2}{m} \log(U-L))$ [40], where m is the total number of given separation gaps, and $n = |V|$.

3.3. Axiomatic Distance Between Incomplete Ratings (Possibly with Ties)

Defining a penalty function on separation gaps is equivalent to quantifying the distance between them. Cook and Kress [13] proposed a distance between complete ratings. This distance function was generalized to incomplete ratings in Fishbain and Moreno-Centeno [11]. This generalized distance called the *normalized projected Cook-Kress distance* (NPCK) uniquely satisfies a set of desirable metric-like axioms. Given incomplete ratings $\boldsymbol{a}^1$ and $\boldsymbol{a}^2$, the NPCK distance between the implied separation gaps is defined as

$$d_{NPCK}(\boldsymbol{a}^1, \boldsymbol{a}^2) = \mathcal{C}^{1,2} \sum_{i \in V_{\boldsymbol{a}}^1 \cap V_{\boldsymbol{a}}^2} \sum_{j \in V_{\boldsymbol{a}}^1 \cap V_{\boldsymbol{a}}^2} \left| p_{ij}^1 - p_{ij}^2 \right|, \tag{6}$$

where

$$\mathcal{C}^{1,2} = \left(4R \cdot \left\lceil \frac{\left| V_{\boldsymbol{a}}^1 \cap V_{\boldsymbol{a}}^2 \right|}{2} \right\rceil \cdot \left\lfloor \frac{\left| V_{\boldsymbol{a}}^1 \cap V_{\boldsymbol{a}}^2 \right|}{2} \right\rfloor \right)^{-1}, \tag{7}$$

and where $\lfloor \cdot \rfloor$ and $\lceil \cdot \rceil$ represent the floor and ceiling functions, respectively. In the above expression, $\mathcal{C}^{1,2}$ is a normalization constant that guarantees that $0 \leqslant d_{NPCK}(\boldsymbol{a}^1, \boldsymbol{a}^2) \leqslant 1$, and $R := U - L$ is the ratings' range. Since $|p_{ij}^1 - p_{ij}^2| = |p_{ji}^1 - p_{ji}^2|$, eq. (6) can be expressed equivalently as

$$d_{NPCK}(\boldsymbol{a}^1, \boldsymbol{a}^2) = 2\mathcal{C}^{1,2} \sum_{(i,j) \in \mathcal{A}^1 \bigcap \mathcal{A}^2} |p_{ij}^1 - p_{ij}^2|. \tag{8}$$

In Fishbain and Moreno-Centeno [11] the *aggregate rating*, $\boldsymbol{x}^*$, is the optimal solution to the *Ratings Cardinal Aggregation* problem:

$$(\mathrm{R-CA}) \qquad \min_{\boldsymbol{x}} \sum_{k=1}^{m} d_{NPCK}(\boldsymbol{a}^k, \boldsymbol{x}). \tag{9}$$

CA is a special case of the SD model; therefore, it is solvable in polynomial time.

We note that $d_{NPCK}(\boldsymbol{a}^1, \boldsymbol{a}^2) = 0$ and $d_{NPCK}(\boldsymbol{a}^1, \boldsymbol{a}^2) = 1$ indicate that there is total agreement and total disagreement, respectively, between the ratings $\boldsymbol{a}^1$ and $\boldsymbol{a}^2$. The normalization is necessary for the distances in problem (9) to be comparable to each other even when the individuals score different numbers of objects. The normalization constant $\mathcal{C}^{1,2}$ was chosen to address the following difficulties: (i) The numbers of objects rated by each incomplete rating may be different; therefore, the distances in problem (9) are over different dimensional spaces. (ii) Each of the distance calculations in problem (9) is between a complete rating, $\boldsymbol{x}^*$, and an incomplete rating, $\boldsymbol{a}^k$—meaning, it only considers the number of objects rated by the incomplete rating. (iii) Distances in higher dimensional spaces tend to be considerably larger than distances in lower-dimensional spaces—from eq. (6), observe that the number of summands of each distance term $d_{NPCK}(\boldsymbol{a}^k, \boldsymbol{x})$ in problem (9) is squarely proportional to the size of $V_{\boldsymbol{a}}^k$.

3.4. Axiomatic Distance Between Incomplete Rankings (Possibly with Ties)

Kemeny and Snell [12] proposed a distance between complete rankings. This distance function was generalized to incomplete rankings in Moreno-Centeno and Escobedo [26]. The authors proved that this generalization, called the *normalized projected Kemeny-Snell distance* (NPKS), uniquely satisfies a set of desirable metric-like axioms. Given incomplete rankings $\boldsymbol{b}^1$ and $\boldsymbol{b}^2$, the NPKS distance is defined as

$$\begin{aligned} d_{NPKS}(\boldsymbol{b}^1, \boldsymbol{b}^2) &= \mathcal{D}^{1,2} \sum_{i \in V_{\boldsymbol{b}}^1 \bigcap V_{\boldsymbol{b}}^2} \sum_{j \in V_{\boldsymbol{b}}^1 \bigcap V_{\boldsymbol{b}}^2} \tfrac{1}{4} \left| \operatorname{sign}(b_i^1 - b_j^1) - \operatorname{sign}(b_i^2 - b_j^2) \right| \\ &= \mathcal{D}^{1,2} \sum_{(i,j) \in \mathcal{B}^1 \bigcap \mathcal{B}^2} \tfrac{1}{2} \left| \operatorname{sign}(b_i^1 - b_j^1) - \operatorname{sign}(b_i^2 - b_j^2) \right| \end{aligned} \tag{10}$$

(since $|\text{sign}(b_i^1 - b_j^1) - \text{sign}(b_i^2 - b_j^2)| = |\text{sign}(b_j^1 - b_i^1) - \text{sign}(b_j^2 - b_i^2)|$ for all i,j), where

$$\mathcal{D}^{1,2} = \left[\frac{\left|V_{\boldsymbol{b}}^1 \cap V_{\boldsymbol{b}}^2\right| \cdot \left(\left|V_{\boldsymbol{b}}^1 \cap V_{\boldsymbol{b}}^2\right| - 1\right)}{2}\right]^{-1}. \tag{11}$$

The eq. (10) summand is the Kemeny and Snell [12] distance function term associated with the ranks given to objects i and j by $\boldsymbol{b}^1$ and $\boldsymbol{b}^2$; $\mathcal{D}^{1,2}$ is a normalization constant that guarantees that $0 \leqslant d_{NPKS}(\boldsymbol{b}^1, \boldsymbol{b}^2) \leqslant 1$. The distance $d_{NPKS}(\boldsymbol{b}^1, \boldsymbol{b}^2)$ has the following natural interpretation: The distance between two incomplete rankings is proportional to the number of *rank reversals* between them. A rank reversal is incurred whenever two objects have a different relative order in the rankings $\boldsymbol{b}^1$ and $\boldsymbol{b}^2$. Similarly, a *half rank reversal* is incurred whenever two objects are tied in one ranking but not in the other ranking. In Moreno-Centeno and Escobedo [26] the *aggregate ranking*, $\boldsymbol{y}^*$, is the optimal solution to the *Rankings Ordinal Aggregation* problem:

$$(\text{R} - \text{OA}) \qquad \min_{\boldsymbol{y}} \sum_{k=1}^{m} d_{NPKS}(\boldsymbol{b}^k, \boldsymbol{y}). \tag{12}$$

Problem R-OA is NP-hard whether the input rankings are complete [15] or incomplete [26].

We note that when there is total agreement between $\boldsymbol{b}^1$ and $\boldsymbol{b}^2$ in the ordinal positions of the objects ranked in common, $d_{NPKS}(\boldsymbol{b}^1, \boldsymbol{b}^2)$ is equal to 0; when there is total disagreement, their distance is equal to 1; otherwise, the distance is strictly between 0 and 1 and proportional to the level of disagreement. The normalization is necessary for the distances in problem (12) to be comparable to each other even when the individuals rank different numbers of objects. The normalization constant $\mathcal{D}^{1,2}$ was chosen to address an analog set of difficulties as $\mathcal{C}^{1,2}$ for incomplete ranking aggregation.

4. Logical Couplings for Multimodal Aggregation

This work develops mathematical models for the joint aggregation of a set of cardinal evaluations $\{\boldsymbol{a}^k\}_{k=1}^m$ and a set of ordinal evaluations $\{\boldsymbol{b}^k\}_{k=1}^m$. The proposed consensus aggregation models are designed to find a cardinal-ordinal evaluation that least disagrees with the multimodal inputs, quantified through an appropriate pair of (axiomatic) distances. These models can be compactly written as

$$(\text{COA}) \qquad \min_{\boldsymbol{x},\boldsymbol{y}} \sum_{k=1}^{m} w_C^k \, d_C(\boldsymbol{a}^k, \boldsymbol{x}) + \sum_{k=1}^{m} w_O^k \, d_O\left(\boldsymbol{b}^k, \boldsymbol{y}\right), \tag{13}$$

where, respectively, $d_C(\cdot,\cdot), d_O(\cdot,\cdot)$ denote unspecified ordinal and cardinal distance functions; parameters w_C^k, w_O^k denote weights assigned to the cardinal and ordinal information from judge k; and variable vectors $\boldsymbol{x},\boldsymbol{y}$ denote the aggregate cardinal and ordinal evaluations. The full contents of these models—objective function, constraints, auxiliary variables—depend on the choices of distance function and the aggregate-evaluation domains (e.g., complete, with ties, etc.). They also depend on the requisite logic to be enforced (e.g., bounds on the solution values, linearized expressions, etc.). The following subsections introduce modeling components to *couple* or logically interrelate $\boldsymbol{x}$ and $\boldsymbol{y}$. In particular, they consider three general interrelationships between cardinal and ordinal evaluations: rating and ranking preferences, numerical and ordinal estimates, and rating and approval preferences.

Before proceeding, it is worthwhile to explain that, while some researchers have defined analogous versions of the COA objective function given by (13) to aggregate multimodal evaluations (e.g. [35–39]), this is where the fundamental similarities with the proposed approach stop. The full specification of the optimization models is starkly different. Specifically, these alternative approaches effectively optimize from the perspective of only one solution modality (e.g., priority vectors), which often differs from the input modalities. Additionally, their solution vector values are compared to the multimodal inputs via non-axiomatic distance functions that directly mix different modalities of data. Conversely, the proposed approach determines the optimal solution using a multimodal vector that matches the input modalities, and it utilizes distance functions that are specific to each modality. Another key difference from the proposed approach is that these alternative methods do not attempt to logically interrelate cardinal and ordinal evaluations from an axiomatic basis. We refer the reader to Section 2 for more details on alternative approaches for utilizing multimodal data.

4.1. Coupling Rating and Ranking Preferences

For the decision-making context, an aggregate rating (i.e., cardinal evaluation) $\boldsymbol{x}$ and an aggregate ranking (i.e., ordinal evaluation) $\boldsymbol{y}$ are coupled by requiring that $\boldsymbol{y} = \mathsf{rank}(\boldsymbol{x})$. This coupling guarantees that objects with higher rating values in the consensus solution also obtain better ranking positions. The following theorem demonstrates how to enforce this logic by adding $O(n^2)$ linear constraints and auxiliary binary variables.

Theorem 1 (*Rating-Ranking Coupling*). Let $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{Z}^n$ be complete rating and ranking vectors that allow ties. Then, the following constraints ensure the preferences expressed in $\boldsymbol{x}$ and $\boldsymbol{y}$ are logically interrelated, i.e., that objects with higher cardinal scores are also ranked higher (receive lower ordinal values).

$$y_i + \sum_{j \neq i} z_{ij} = n \qquad i = 1, .., n, \tag{14a}$$

$$x_i - x_j + 1 \leqslant M_1 \, z_{ij} \qquad i, j = 1, .., n; i \neq j, \tag{14b}$$

$$-x_i + x_j \leqslant M_2 (1 - z_{ij}) \qquad i, j = 1, .., n; i \neq j, \tag{14c}$$

$$z_{ij} \in \{0, 1\} \qquad i, j = 1, .., n; i \neq j. \tag{14d}$$

where $\boldsymbol{z} \in \{0, 1\}^{n^2}$ are auxiliary variables and M_1, M_2 are constants large enough so that constraint (14b) is satisfied whenever $z_{ij} = 1$ and constraint (14c) is satisfied whenever $z_{ij} = 0$, for any feasible setting of $\boldsymbol{x}$.

Proof. We give a preference interpretation to auxiliary variable z_{ij}:

$$z_{ij} = \begin{cases} 1 & \text{if object } i \text{ is preferred or tied with object } j, \\ 0 & \text{otherwise.} \end{cases} \tag{15}$$

Constraint (14a) provides a one-to-one relationship between variables z_{ij} and the ranking position of object i, according to the convention for expressing rankings with ties described in Section 3.1. That is, tallying the number of other objects over which i is preferred or tied, given by $\sum_{j \neq i} z_{ij}$, and subtracting this total from n provides the object's rank, y_i. Next, constraints (14b) and (14c) enforce that $\boldsymbol{y} = \mathsf{rank}(\boldsymbol{x})$, by considering the implications of each preference modality onto the other.

First, the ordinal preferences implied by the relationships between aggregate cardinal variables x_i and x_j are encapsulated with two cases:

- **Cardinal to Ordinal Preferences, Case 1**: $x_i \geqslant x_j$.
  The left-hand side of constraint (14b) is positive, which forces $z_{ij} = 1$, i.e., i is tied or ranked better than j. This setting makes the right-hand side of constraint (14c) equal to 0; the constraint is automatically satisfied since $x_j - x_i \leq 0$.
- **Cardinal to Ordinal Preferences, Case 2**: $x_i < x_j$. The left-hand side of constraint (14c) is positive, which forces $z_{ij} = 0$, i.e., i is ranked worse than j. This setting makes the right-hand side of constraint (14b) equal to 0; the constraint is automatically satisfied since $x_i - x_j < 0$.

Second, the cardinal preferences implied by the aggregate ordinal preferences between i and j, represented by z_{ij}, are encapsulated with two cases:

- **Ordinal to Cardinal Preferences, Case 1**: $z_{ij} = 0$.
  The right-hand side of constraint (14b) is 0, which implies that $x_i + 1 \leqslant x_j$ (note that (14c) becomes redundant). In other words, if i receives a worse rank than j, then it must also receive a lower rating.
- **Ordinal to Cardinal Preferences, Case 2**: $z_{ij} = 1$. The right-hand side of constraint (14c) is 0, which implies that $x_j \leqslant x_j$. In other words, if i is tied or ranked better than j, then it must receive at least as high a rating as the latter. □

A couple of clarifications are in order. First, the assumption in Theorem 1 that the aggregate rating vector is integral is without loss of generality since it is possible to specify any desired rating precision through the interpretation of $\boldsymbol{x}$. Expressly, bounding x_i as $L/\mu \leqslant x_i \leqslant U/\mu$, for $i = 1, \ldots, n$, where L and U are the rating bounds and $\mu = 1/p$ is the desired score precision or minimum rating separation, with an integer $p \geq 1$, x_i can be interpreted as the number of minimum separation gaps from L obtained by object i (see Section 5 for more details). Given this interpretation, the tightest possible "Big-M" constants for constraints (14b) and (14c) are $M_1 = (U - L + 1)/\mu$ and $M_2 = (U - L)/\mu$, respectively. Second, notice that constraints (14a)–(14d) are sufficient to prevent cycles in the aggregate ordinal preferences because the consensus ranking positions of any three objects $h, i, j \in V$ are directly implied by the ordering of their consensus rating values (each of which cannot assume more than one cardinal value).

4.2. Coupling Cardinal and Ordinal Estimates

Although axiomatic aggregation methods are traditionally associated with social (i.e., human) contexts, their use extends to numerous other situations requiring the aggregation of conflicting information from non–human sources. For example, consensus aggregation has been widely used in information retrieval to derive representative lists of relevant documents in databases and perform metasearch [10] and in bioinformatics to build genetic maps and consolidate gene expression results [9]. Akin to the social context, consensus aggregation methods are used in these settings to consolidate heterogeneous evaluations that may be inconsistent, unreliable, and/or biased. The proposed multimodal consensus aggregation may apply to such contexts where cardinal and ordinal assessments are available. However, an important distinction is that there may be multiple options for coupling $\boldsymbol{x}$ and $\boldsymbol{y}$, the appropriateness of which must be judged from the context at hand. The requirement that objects that receive higher cardinal values must also receive lower ordinal values was covered in

Section 4.1. This subsection covers another coupling relevant for estimating some objectively quantifiable characteristics, namely the requirement that objects that receive higher cardinal values must also receive higher ordinal values.

Theorem 2 (*Cardinal and Ordinal Estimate Coupling*). Let $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{Z}^n$ be complete cardinal and ordinal vectors that allow ties. The following mixed integer linear constraints interrelate $\boldsymbol{x}$ and $\boldsymbol{y}$ so that objects that receive higher cardinal values also receive higher ordinal values:

$$y_i + \sum_{j \neq i} z_{ij} = n \qquad i = 1, .., n, \tag{16a}$$
$$x_i - x_j + 1 \leqslant M_1(1 - z_{ij}) \qquad i, j = 1, .., n; i \neq j, \tag{16b}$$
$$-x_i + x_j \leqslant M_2\, z_{ij} \qquad i, j = 1, .., n; i \neq j, \tag{16c}$$
$$z_{ij} \in \{0, 1\} \qquad i, j = 1, .., n; i \neq j. \tag{16d}$$

where $\boldsymbol{z} \in \{0, 1\}^{n^2}$ are auxiliary variables and M_1, M_2 are constants large enough so that constraint (16b) is always satisfied when $z_{ij} = 0$ and constraint (16c) is always satisfied when $z_{ij} = 1$, for any feasible setting of $\boldsymbol{x}$.

Proof. We give a slightly different interpretation to auxiliary variable z_{ij}:

$$z_{ij} = \begin{cases} 1 & \text{if object } i \text{ exhibits a lower quantity of the} \\ & \text{observed characteristic than object } j, \\ 0 & \text{otherwise.} \end{cases} \tag{17}$$

The remainder of the proof uses similar logic as that of Theorem 1 and is omitted for brevity. □.

4.3. Coupling Rating and Approval Preferences

Approval voting is a relatively simple ordinal voting method that has received recent attention from various communities [6]. The technique seeks to divide the objects into an "approved" (i.e., winning) subset and a "disapproved" (i.e., losing) subset. An *approval ballot* is defined as a subset $V^k_+ \subseteq V$ that indicates those objects judge k deems as approved; all other objects evaluated by judge k belong to a disapproved subset $V^k_- \subseteq V \setminus V^k_+$. Judge k's approval ballot can be equivalently expressed using a binary (ordinal) vector $\boldsymbol{b}^k$ defined as

$$b^k_i = \begin{cases} 1 & \text{if object } i \text{ is approved,} \\ 0 & \text{if object } i \text{ is disapproved,} \\ \bullet & \text{if object } i \text{ is not evaluated by judge } k. \end{cases} \tag{18}$$

An aggregate approval voting vector $\boldsymbol{y} \in \{0, 1\}^n$ is defined similarly, with the only difference that the last case from eq. (18) is unnecessary since the solution vector is assumed to be complete. Approval voting has been criticized for its oversimplification of the collective preferences. Here, we propose logical expressions that can be used to construct ratings and approval-voting joint aggregation models.

Before proceeding, it is pertinent to mention a model recently introduced by Dong et al. [41] for coupling ranking and approval preferences in group decision-making. The model assumes that each judge submits a weak ordering, along with a dividing line separating the judge's approved and disapproved alternatives. This preference data structure is complemented with an axiomatic ranking-approval distance function, which is proved to be unique. Notice, however, that the two types of preferences coupled in Dong et al. [41] are ordinal, whereas the proposed methodology logically interrelates ordinal and cardinal modalities. Specifically, the following theorem couples the collective approval voting vector $\boldsymbol{y}$ with the collective rating vector $\boldsymbol{x}$ through a set of mixed-integer linear constraints.

Theorem 3 (*Rating and Approval Aggregation*). *Let $\boldsymbol{x} \in \mathbb{Z}^n$ be a complete rating that allows ties and $\boldsymbol{y} \in \{0, 1\}^n$ be a complete approval voting vector. The addition of the following mixed integer linear constraints couples $\boldsymbol{x}$ and $\boldsymbol{y}$ so that approved objects in the aggregate evaluation receive higher cardinal values than disapproved objects:*

$$x_i - x_j + 1 \leqslant M_1 z_{ij} \qquad i, j = 1, \ldots, n; i \neq j, \tag{19a}$$
$$y_i - y_j \leqslant z_{ij} \qquad i, j = 1, \ldots, n; i \neq j, \tag{19b}$$
$$-y_i + y_j \leqslant (1 - z_{ij}) \qquad i, j = 1, \ldots, n; i \neq j, \tag{19c}$$
$$z_{ij} \in \{0, 1\} \qquad i, j = 1, \ldots, n; i \neq j. \tag{19d}$$

Proof. Constraints (19a)-(19d) reflect the required coupling logic through the use of auxiliary variables z_{ij}, which are interpreted as in eq. (15). To demonstrate this, we evaluate the implications of $\boldsymbol{x}$ on $\boldsymbol{y}$ and then the implications of $\boldsymbol{y}$ on $\boldsymbol{x}$. First, the approval preferences implied by the relationships between aggregate cardinal variables x_i and x_j is encapsulated with two cases:

- **Cardinal to Approval, Case 1**: $x_i \geqslant x_j$.
  The left-hand side of constraint (19a) is positive, which forces $z_{ij} = 1$. In turn, this setting makes the right-hand side of constraint (19c) equal to 0, which is equivalent to requiring that $y_j \leqslant y_i$ (note that constraint (19b) becomes redundant). In other words, when object i receives a rating value that is at least as good as the rating object j receives, it is not possible simultaneously for i to be disapproved and j to be approved, i.e., we cannot have that $y_j > y_i$.
- **Cardinal to Approval, Case 2**: $x_i < x_j$. The left-hand side of (19a) is non-positive and, therefore, z_{ij} is allowed to assume any value, i.e., no coupling with y_i and y_j is enforced.

Second, the cardinal preferences implied by the relationships between aggregate approval voting variables y_i and y_j is encapsulated with three cases:

- **Approval to Cardinal, Case 1**: $y_i = y_j$.
  The left-hand sides of constraints (19b) and (19c) are 0 and, therefore, z_{ij} is allowed to assume any value, i.e., no coupling with x_i and x_j is enforced.
- **Approval to Cardinal, Case 2**: $y_i = 0, y_j = 1$.
  Constraint (19c) implies that $z_{ij} = 0$. This in turn implies that $x_i + 1 \leqslant x_j$ in constraint (19a). In other words, if j is approved and i is disapproved in the aggregate approval preferences, it must also be the case that j receives a higher cardinal value than i.
- **Approval to Cardinal Case 3**: $y_i = 1, y_j = 0$. Constraint (19b) implies that $z_{ij} = 1$, which makes constraint (19a) redundant. □

It is important to remark that, because constraints are generated for all $i, j \in \{1, \ldots, n\}$, Case 3 forces $z_{ji} = 0$ in the respective constraints where the labels i and j are exchanged (i.e., this is enforced by Case 2 after the exchange).

By adding constraints (19a)-(19d) to problem (13), it is possible to find a rating-approval consensus using a suitable pair of distances. Example distances used to aggregate ratings include d_{NPCK} (discussed in Section 3.3). Example distances used to aggregate approval ballots include the Hamming distance [42].

5. Axiomatic Distance-based Rating and Ranking Aggregation

The previous section derived linear expressions for logically interrelating different types of cardinal and ordinal evaluations. The respective constraint sets are only one part of the mathematical modeling components needed to obtain an explicit representation of the consensus aggregation problem (see (13)) for a specific cardinal-ordinal distance pair. To complete the optimization model, it is necessary to include the corresponding objective function expressions and other specialized constraints and auxiliary variables. The remainder of this paper focuses on a multimodal consensus aggregation model that combines the rating aggregation problem via distance d_{NPCK} (problem (9), denoted as R-CA) and the ordinal aggregation problem via distance d_{NPKS} (problem (12), denoted as R-OA). We denote this as the *Ratings and Rankings Cardinal and Ordinal Aggregation* problem, or RR-COA, for short.

This section is organized as follows. First, SubSection 5.1 introduces an abbreviated version of RR-COA and demonstrates that the problem is NP-hard. Second, SubSection 5.2 derives an exact mixed-integer linear program (MILP) reformulation to solve this problem exactly. It also proposes a strengthened version of the formulation that incorporates structural valid inequalities. Third, SubSection 5.3 derives a convex relaxation, which serves as an efficient and effective heuristic capable of solving instances with a very large number of objects. Lastly, SubSection 5.4 describes supplementary techniques for analyzing the RR-COA solution.

5.1. The Ratings and Rankings Cardinal and Ordinal Aggregation Problem

RR-COA can be written in abbreviated form as:

$$(\mathrm{RR-COA}) \qquad \min_{\boldsymbol{x},\boldsymbol{y}} \sum_{k=1}^{m} d_{NPCK}(\boldsymbol{a}^{\mathbf{k}}, \boldsymbol{x}) + \sum_{k=1}^{m} d_{NPKS}\left(\boldsymbol{b}^{\mathbf{k}}, \boldsymbol{y}\right), \tag{20a}$$

$$\text{subject to} \qquad \boldsymbol{y} = \mathrm{rank}(\boldsymbol{x}), \tag{20b}$$

$$0 \leqslant x_i \leqslant \frac{U-L}{\mu} \qquad i = 1, .., n, \tag{20c}$$

$$x_i, y_i \in \mathbb{Z} \qquad i = 1, .., n. \tag{20d}$$

It is important to elaborate on the fundamental assumptions inherent in this formulation. The goal of this model is to give fair representation/weight to each of the evaluation inputs and each data modality; this is enforced by giving equal weight to the cumulative d_{NPCK} distance term and the cumulative d_{NPKS} distance term. If justified by a particular context, different weight parameters can be assigned to the two cumulative distance terms and/or to their individual summands (see (13)).

Additionally, the parameter $\mu = 1/p$ specifies the *score precision* or *minimum separation gap* in the solution's rating values of non-tied objects, where $p \in \mathbb{Z}_+$. Accordingly, x_i gives the number of minimum separation gaps from the lowest rating value, L, of object i. That is, the consensus rating value scaled according to the original rating range is obtained via the expression $L + \mu x_i$.

Higher solution precision than is specified in the rating inputs may be needed to incorporate enough separation gaps in the aggregate rating. As a basic requirement, the value of μ must not implicitly exclude any possible ordering of n objects to be feasible for RR-COA (recall that the ranking solution is induced by ordering the $\boldsymbol{x}$ values in non-increasing order). For example, using a scoring range of 0.0 to 10.0 and minimum precision of $\mu = 0.5$, it is impossible to obtain a strict ranking of more than 21 objects. Therefore, the value of μ should be set at most to $(U - L)/(n - 1)$. Furthermore, μ should be small enough to capture large intensities of preference transitively implied through multiple pairwise comparisons. Continuing with the example, if a judge evaluates only objects h and i and gives them scores of 10.0 and 0.0 (a difference of 20 minimum-separation gaps), respectively, and another judge evaluates only objects i and j and gives them scores of 10.0 and 0.0, respectively, the combination of their scores would suggest a stronger intensity of preference between objects h and j than between the two explicitly compared pairs (a difference of 40 minimum-separation gaps). Therefore, the collective preferences over n objects may transitively imply as many as $[(n-1)(U-L)]/\mu_{input}$ separation gaps between two objects, where μ_{input} is the minimum allowable precision in the input ratings. Therefore, the value of μ should be at least $\mu_{input}/[(n-1)(U-L)]$. However, to enhance solution interpretability, one can first solve with μ set to its lower bound and then increase μ (and thus increase interpretability) and re-solve. This process can be repeated as long as the number of gaps separating every pair of objects in the solution remains constant between runs and μ does not exceed its upper bound.

It is also worth highlighting that, while the formulation enforces the rating and ranking coupling featured in Section 4.1 via constraint (20b), a different logical relationship between the aggregate cardinal and ordinal evaluations can be used. For instance, [43] recently applied a modified version of RR-COA (using one of the MILPs developed herein) in the context of crowdsourced computation; this field studies how to combine the abilities of multiple humans to complete complex tasks. The authors enforced the coupling for cardinal and ordinal estimates introduced in Section 4.2 to perform two related but distinct crowdsourced computation tasks: ordering a set of images based on the number of dots they contain (fewest to most) and estimating the number of dots each image contains. The results therein attest that eliciting and aggregating multimodal information can improve the quality of crowdsourced estimates.

Since d_{NPCK} and d_{NPKS} are generalized versions of the Cook and Kress [13] complete rating distance and the Kemeny and Snell [12] complete ranking distance, respectively, problem (20) can be used to solve the complete ranking and rating aggregation problem, also previously undefined in the literature. Next, we establish that RR-COA is NP-hard by reducing it from R-OA (problem (12)), which is NP-hard [15].

Lemma 1. Problem RR-COA is NP-hard.

Proof. Given an instance of R-OA (a set of incomplete rankings $\{\boldsymbol{b}^k\}_{k=1}^m$), one can transform it in polynomial time to an instance of RR-COA as follows. Keep $\{\boldsymbol{b}^k\}_{k=1}^m$ unchanged and create a set of ratings $\{\boldsymbol{a}^k\}_{k=1}^m$ such that each rating evaluates exactly one object (the choice of object is irrelevant; in fact, all of the ratings can evaluate the same object). From the definition of d_{NPCK} (eq. (6)), it follows that, for every $\boldsymbol{x}$, the first summand in RR-COA will be equal to 0. Therefore, with this choice of ratings, the optimal solution to RR-COA will be $\boldsymbol{y}^*$, that is, the optimal solution to R-OA.□

5.2. Deriving an Exact MILP Formulation of RR-COA

The objective functions of R-CA and R-OA are nonlinear. This subsection linearizes and combines both objectives to construct an exact mixed-integer linear programming (MILP) formulation of RR-COA. It is helpful to begin with R-OA and to define parameters $\hat{b}_{ij}^k$ as

$$\hat{b}_{ij}^k = \begin{cases} 1 & \text{if } b_i^k \leqslant b_j^k, \\ -1 & \text{if } b_i^k > b_j^k, \\ 0 & \text{if } i = j, \end{cases} \tag{21}$$

for $(i,j) \in \mathcal{B}^k$ and $k \in \{1, \ldots, m\}$. The R-OA solution can be expressed as:

$$\begin{aligned}
&\underset{\boldsymbol{y}}{\operatorname{argmin}} \sum_{k=1}^{m} d_{NPKS}\left(\boldsymbol{b}^{\mathbf{k}}, \boldsymbol{y}\right) \\
&= \underset{\boldsymbol{y}}{\operatorname{argmax}} -2\left[\sum_{k=1}^{m} d_{NPKS}\left(\boldsymbol{b}^{\mathbf{k}}, \boldsymbol{y}\right)\right] + m
\end{aligned} \tag{22a}$$

$$= \underset{\boldsymbol{y}}{\operatorname{argmax}} \sum_{k=1}^{m} 1 - 2d_{NPKS}\left(\boldsymbol{b}^{k}, \boldsymbol{y}\right) \tag{22b}$$

$$= \underset{\boldsymbol{z}}{\operatorname{argmax}} \sum_{k=1}^{m} \mathcal{D}^{k} \sum_{(i,j)\in\mathcal{B}^{k}} \hat{b}_{ij}^{k} z_{ij}, \tag{22c}$$

where the latter equation applies an equivalent representation of distance d_{NPKS} derived in [27]. Expressly, the resulting maximization problem linearizes (10) by introducing parameters $\hat{b}_{ij}^{k}$, defined by (21), and binary variables $z_{ij} \in \{0,1\}$, where $i,j \in V$ and $k \in \{1,\ldots,m\}$ (these auxiliary variables can interpreted as in (15)). To yield the corresponding aggregate ranking, the equation $y_i + \sum_{j\neq i} z_{ij} = n$ is solved, for all i (see Section 4.1 for more details). Next, the solution to R-CA can be re-expressed as:

$$\begin{aligned}
&\underset{\boldsymbol{x}}{\operatorname{argmin}} \sum_{k=1}^{m} 2C^{k} \sum_{(i,j)\in\mathcal{A}^{k}} \left|\mu(x_i - x_j) - p_{ij}^{k}\right| \\
&= \underset{\boldsymbol{x}}{\operatorname{argmax}} -2\left[\sum_{k=1}^{m} 2C^{k} \sum_{(i,j)\in\mathcal{A}^{k}} \left|\mu(x_i - x_j) - p_{ij}^{k}\right|\right]
\end{aligned} \tag{23a}$$

$$= \underset{\boldsymbol{t}}{\operatorname{argmax}} \sum_{k=1}^{m} -4C^{k} \sum_{(i,j)\in\mathcal{A}^{k}} t_{ij}^{k}, \tag{23b}$$

where auxiliary variables $t_{ij}^{k} \geq 0$ are used to substitute the respective absolute-value terms by requiring equivalently that $t_{ij}^{k} \geqslant \mu(x_i - x_j) - p_{ij}^{k}$ and $t_{ij}^{k} \geqslant -\mu(x_i - x_j) + p_{ij}^{k}$, for $(i,j) \in \mathcal{A}^{k}$ and $k \in \{1,\ldots,m\}$. Recall that the d_{NPCK} and d_{NPKS} normalization constants are indexed above only by a single index in contrast to their two-index definitions (see (7) and (11)) because in the consensus aggregation problem each input rating/ranking is always compared to a complete rating/ranking (the aggregate evaluation).

From the above derivations and the results of Section 4.1, the *Base RR-COA MILP* is as follows:

$$\underset{\boldsymbol{t},\boldsymbol{x},\boldsymbol{z}}{\operatorname{argmax}} \sum_{k=1}^{m} -4C^{k} \sum_{(i,j)\in\mathcal{A}^{k}} t_{ij}^{k} + \sum_{k=1}^{m} \mathcal{D}^{k} \sum_{(i,j)\in\mathcal{B}^{k}} \hat{b}_{ij}^{k} z_{ij}, \tag{24a}$$

subject to

$$t_{ij}^{k} - \mu(x_i - x_j) \geqslant -p_{ij}^{k} \qquad (i,j) \in \mathcal{A}^{k}, k = 1,\ldots,m, \tag{24b}$$

$$t_{ij}^{k} + \mu(x_i - x_j) \geqslant p_{ij}^{k} \qquad (i,j) \in \mathcal{A}^{k}, k = 1,\ldots,m, \tag{24c}$$

$$x_i - x_j \leqslant M_1 z_{ij} - 1 \qquad i,j = 1,\ldots,n; i \neq j, \tag{24d}$$

$$-x_i + x_j \leqslant M_2(1 - z_{ij}) \qquad i,j = 1,\ldots,n; i \neq j, \tag{24e}$$

$$x_i \leqslant \frac{U - L}{\mu} \qquad i = 1,\ldots,n, \tag{24f}$$

$$z_{ij} \in \{0,1\} \qquad i,j = 1,\ldots,n; i \neq j, \tag{24g}$$

$$x_i \in \mathbb{Z}^{+}_{\cup\{\mathbb{0}\}} \qquad i = 1,\ldots,n. \tag{24h}$$

Additionally, we seek to enhance the computational performance of RR-COA MILP through the incorporation of structural valid inequalities (VIs). The insight behind the VIs is linked with the preference relations that are guaranteed by pairs and triplets of variables z_{ij}. More specifically, the following linear expressions are satisfied by any set of values z_{ij} that induces a complete non-strict ranking of n objects [44]:

$$z_{ij} + z_{ji} \geqslant 1 \qquad i,j = 1,\ldots,n; i \neq j \tag{25a}$$

$$z_{ij} - z_{kj} - z_{ik} \geqslant -1 \qquad i,j,k = 1,\ldots,n; i \neq j \neq k \neq i. \tag{25b}$$

In short, these expressions enforce the properties of a weak ordering (a binary relation that is reflexive, transitive, and total). We denote the resulting formulation as the *Enhanced RR-COA MILP* and evaluate its comparative performance with the Base RR-COA MILP in Section 6. It is worth adding that (25a) and (25b) are logically equivalent expressions of two of the three members of the *basic family* of facet defining inequalities of the weak order polytope (the convex hull of the

characteristic vectors induced by all weak orders on n objects); the third member is the upper bound constraint $z_{ij} \leq 1$ for $i,j = 1,\ldots,n$. This family of VIs represents only a subset of all facet defining inequalities of the polytope known to date (e.g. [45]).

5.3. Convex Relaxation of RR-COA

It is useful to return to the original (nonlinear, nonconvex) formulation of RR-COA, which can be written as:

$$\underset{\boldsymbol{x}}{\text{argmin}} \sum_{k=1}^{m} \left[2C^k \sum_{(i,j)\in\mathcal{A}^k} \left| \mu(x_i - x_j) - p_{ij}^k \right| \right] + \sum_{k=1}^{m} \left[\frac{1}{2} \mathcal{D}^k \sum_{(i,j)\in\mathcal{B}^k} \left| \text{sign}(x_i - x_j) - \text{sign}(b_j^k - b_i^k) \right| \right], \tag{26a}$$

$$\text{subject to} \quad 0 \leqslant x_i \leqslant \frac{U-L}{\mu} \qquad i = 1,\ldots,n, \tag{26b}$$

$$x_i \in \mathbb{Z} \qquad i = 1,\ldots,n. \tag{26c}$$

Notice that the above formulation replaces the ordinal vector $\boldsymbol{y}$ and the cardinal-ordinal coupling expressions in formulation (24) with the terms $\text{sign}(x_i - x_j)$ in the objective function. Additionally, the argument of the other sign function in the objective function is $(b_j^k - b_i^k)$ and not $(b_i^k - b_j^k)$, as in eq. (10). This modified argument matches the rating-ranking coupling discussed in Section 4.1, which requires that higher cardinal numbers are assigned to objects judged as more preferable while higher ordinal numbers are assigned to objects judged as less preferable.

As the preceding subsection demonstrates, each cardinal-aggregation term in (26a) is a convex piecewise-linear term that is easily linearized. By contrast, each ordinal-aggregation term in (26a) is highly nonconvex and nonlinear (see (10)). This subsection proposes to approximate the function $g^k(x_i,x_j) := |\text{sign}(x_i - x_j) - \text{sign}(b_j^k - b_i^k)|$, the k^{th} summand of the second sum in (26a), with a tight upper-convex (piecewise-linear) envelope, $h^k(x_i,x_j)$:

$$h^k(x_i,x_j) = \begin{cases} \max\{0, x_i - x_j + 1\} & \text{if sign } (b_j^k - b_i^k) = -1 \\ \max\{-x_i + x_j, x_i - x_j\} & \text{if sign } (b_j^k - b_i^k) = 0 \\ \max\{-x_i + x_j + 1, 0\} & \text{if sign } b_j^k - b_i^k) = 1. \end{cases} \tag{27}$$

Fig. 1 shows that $h^k(x_i,x_j)$ approximates $g^k(x_i,x_j)$ and provides a tight convex envelope.

Replacing $g^k(x_i,x_j)$ with $h^k(x_i,x_j)$ and linearizing the cardinal-aggregation terms in the objective function yields the following convex relaxation of RR-COA, denoted as c-RR-COA:

$$\underset{\boldsymbol{x},\boldsymbol{t},\boldsymbol{h}}{\text{argmin}} \sum_{k=1}^{m} 2C^k \sum_{(i,j)\in\mathcal{A}^k} t_{ij}^k + \sum_{k=1}^{m} \frac{1}{2} \mathcal{D}^k \sum_{(i,j)\in\mathcal{B}^k} h_{ij}^k, \tag{28a}$$

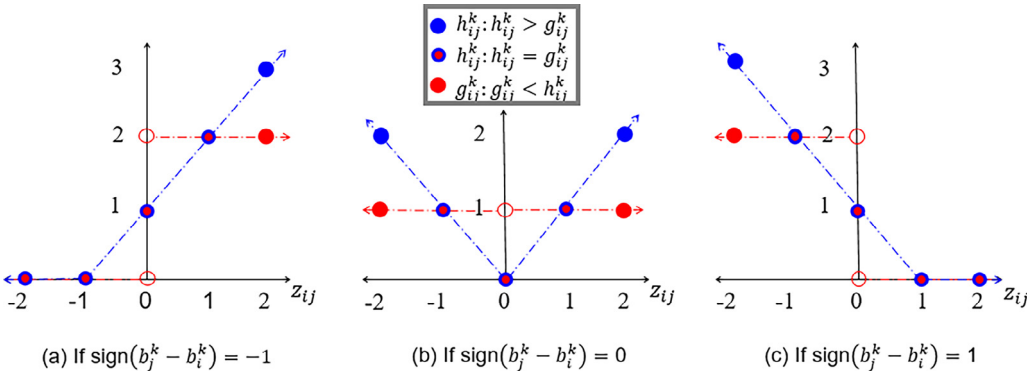

Fig. 1. Relationships between $g^k(x_i,x_j)$ (shorthand g_{ij}^k) and its upper convex envelope $h^k(x_i,x_j)$ (shorthand h_{ij}^k), for each of the three possible values of $\text{sign}(b_j^k - b_i^k)$. For ease of illustration, the domain $(x_i,x_j) \in \mathbb{Z}^2$ is projected into a 1-dimensional space via the auxiliary variable $z_{ij} := x_i - x_j \in \mathbb{Z}^1$.

subject to

$$t_{ij}^k - \mu(x_i - x_j) \geqslant -p_{ij}^k \qquad (i,j) \in \mathcal{A}^k, k = 1,\ldots,m, \tag{28b}$$
$$t_{ij}^k + \mu(x_i - x_j) \geqslant p_{ij}^k \qquad (i,j) \in \mathcal{A}^k, k = 1,\ldots,m, \tag{28c}$$
$$h_{ij}^k - x_i + x_j \geqslant 1 \qquad (i,j) \in \mathcal{B}^k, k = 1,\ldots,m; \text{s.t. } \text{sign}(b_j^k - b_i^k) = -1, \tag{28d}$$
$$h_{ij}^k - x_i + x_j \geq 0 \qquad (i,j) \in \mathcal{B}^k, k = 1,\ldots,m; \text{s.t. } \text{sign}(b_j^k - b_i^k) = 0, \tag{28e}$$
$$h_{ij}^k + x_i - x_j \geq 0 \qquad (i,j) \in \mathcal{B}^k, k = 1,\ldots,m; \text{s.t. } \text{sign}(b_j^k - b_i^k) = 0, \tag{28f}$$
$$h_{ij}^k + x_i - x_j \geq 1 \qquad (i,j) \in \mathcal{B}^k, k = 1,\ldots,m; \text{s.t. } \text{sign}(b_j^k - b_i^k) = 1, \tag{28g}$$
$$h_{ij}^k \geq 0 \qquad (i,j) \in \mathcal{B}^k, k = 1,\ldots,m, \tag{28h}$$
$$t_{ij}^k \geq 0 \qquad (i,j) \in \mathcal{A}^k, k = 1,\ldots,m, \tag{28i}$$
$$0 \leqslant x_i \leqslant \tfrac{U-L}{\mu} \qquad i = 1,\ldots,n, \tag{28j}$$
$$x_i \in \mathbb{Z} \qquad i = 1,\ldots,n. \tag{28k}$$

c-RR-COA is a special case of the convex SD model and, therefore, it is solvable in polynomial time.

5.4. Supplementary Analyses from the RR-COA Solution

Next, we propose a mechanism to identify inconsistencies in the given evaluations (e.g., outliers, judges that are too lenient or too strict, etc.). For instance, this information may be helpful for the lead decision-maker to initiate an investigation of the nature of unusual discrepancies and justify further deliberations (e.g., discussing these inconsistencies with the judges and promoting discussion to alleviate them).

The mechanism uses the rating solution to RR-COA, denoted as $\boldsymbol{x}^{(RR)}$, to identify (i) judges whose evaluations differ the most from the other evaluations and (ii) objects about which judges had particularly divergent evaluations. These judges (objects) are those that assigned (received) scores that disagree the most with $\boldsymbol{x}^{(RR)}$. Specifically, we use the individual contributions to the separation penalty to identify the judges whose evaluations are farthest from $\boldsymbol{x}^{(RR)}$. The contribution of judge $\underline{k} \in \{1,\ldots,m\}$ to the separation penalty is calculated as

$$2C^{\underline{k}} \sum_{(i,j)\in\mathcal{A}^{\underline{k}}} \left|(x_i^{(RR)} - x_j^{(RR)}) - (a_i^{\underline{k}} - a_j^{\underline{k}})\right|. \tag{29}$$

Similarly, we use the separation penalty to identify the objects that engendered remarkably divergent evaluations. These objects are those with the highest contribution to the separation penalty. The contribution of object $\underline{i}$ to the separation penalty is calculated as

$$\sum_{k=1}^{m}\left[\sum_{(\underline{i},j)\in\mathcal{A}^k} 2C^k\left|(x_{\underline{i}}^{(RR)} - x_j^{(RR)}) - (a_{\underline{i}}^k - a_j^k)\right| + \sum_{(j,\underline{i})\in\mathcal{A}^k} 2C^k\left|(x_j^{(RR)} - x_{\underline{i}}^{(RR)}) - (a_j^k - a_{\underline{i}}^k)\right|\right]. \tag{30}$$

6. Computational Tests and Analysis

This section assesses various practical dimensions of the featured aggregation methodology. To this end, it first considers a real-world case study involving the 2007 MSOM Student Paper Competition, referenced succinctly as 2007 MSOM SPC. Afterward, it introduces a procedure for generating synthetic instances motivated by the response styles and other practical considerations observed in the 2007 MSOM SPC. Finally, the synthetic instances allow for a comprehensive computational analysis of the featured methodology. The experiments were performed on machines with 36 GB of RAM shared by two Intel Xeon E5-2680 processors running at 2.40 GHz. The code was written in Python, and the optimization models were solved with CPLEX version 12.9.

6.1. Analysis of 2007 MSOM SPC

This subsection considers the 2007 MSOM SPC case study, consisting of 58 submitted papers and 63 participating judges. Each judge evaluated only three to five papers in the competition, and only three to five judges reviewed each paper. Although the input evaluations are highly incomplete, this instance's paper-judge allocation is considered robust (see the following subsection for more details). It is worthwhile to note that some of the input ratings tended to use the entire (i.e., an expanded) rating scale, others tended to use only the middle scores (i.e., a condensed scale), others tended to use only the top scores (i.e., an optimistic scale), and yet others tended to use only the bottom scores (i.e., a pessimistic scale).

Table 1
Description of evaluation attributes with respective numerical scales.

Attribute	Description	Scale
(A)	Problem importance/interest	1–10
(B)	Problem modeling	0–10
(C)	Analytical results	0–10
(D)	Computational results	0–10
(E)	Paper writing	1–10
(F)	Overall contribution to the field (field contribution, for short)	1–10

Table 2
Numerical score rubric (the journals: MSOM, OR and MS).

Score	Definition/Interpretation
10	Attribute considered is comparable to that of the best papers published in the journals.
8,9	Attribute considered is comparable to that of the average papers published in the journals.
7	Attribute considered is at the minimum level for publication in the journals.
5,6	Attribute considered independently would require a minor revision before publication in the journals.
3,4	Attribute considered independently would require a major revision before publication in the journals.
1,2	Attribute considered would warrant by itself a rejection if the paper were submitted to the journals.
0	Attribute considered is not relevant or applicable to the evaluated paper.

Such diverse response styles align with numerous real-world studies (e.g. [46]). There were also self-contradictions between the input ranking evaluations and the rankings induced by the input rating evaluations of individual judges.

The papers were rated based on six different attributes (see Table 1), gauged according to a precise numerical rubric (see Table 2). The rubric was set according to the respective qualities of papers published in three top-tier domain journals (heretofore referred to as *the journals*): *Manufacturing* & *Service Operations Management* (MSOM), *Operations Research* (OR), and *Management Science* (MS). Each judge also provided an ordinal evaluation (a ranking) of the papers they reviewed (1 = best, 2 = second best, etc.), allowing ties.

Although this precise rubric was provided to the judges, they differed significantly in their evaluations and presumably interpreted the scores differently. Examples of this phenomenon are illustrated for papers 18 and 26 in Tables 3 and 4, respectively. The labels identifying judges and papers have been randomly permuted from their original assignments to preserve the participants' anonymity.

As Table 3 illustrates, in Attributes (B), (C), and (F), paper 18 was given a score of 8 by one judge—i.e., the paper is comparable in problem modeling, analytical results, and field contribution to an average paper in the journals—and scores no greater than 4 by the other four judges—i.e., the paper requires at least a major revision. Such scoring discrepancies are significant and especially pronounced between judges 12 and 42. For the most part, the former considers the paper to be at the average level for publication, while the latter holds the journals should reject it.

A similar discrepancy in subjective judgments involving paper 26 can be seen in Table 4. Therein, judge 14's evaluations do not appear to be on the same scale as the evaluations of the other four judges. On the one hand, in every attribute except (A), judge 14's evaluation indicates that the paper would be rejected. On the other, in all attributes, all other judges deemed the paper worthy of publication—some of them even indicated it would be among the best papers published in the journals!

Such glaring discrepancies in the judges' evaluations over the same papers are commonplace throughout this real-world data set. Henceforth, we use the average scores over the six attributes, excepting scores of 0 (which connote lack of relevance rather than poor quality), as the input ratings of each judge.

Table 5 compares the optimal solutions obtained by the three aggregation models: (i) $\boldsymbol{x}^*$, obtained by aggregating only the ratings via R-CA (Problem (9)); (ii) $\boldsymbol{y}^*$, obtained by aggregating only the rankings via R-OA (problem (12)); and (iii) $\boldsymbol{x}^{(RR)}$ and rank$(\boldsymbol{x}^{RR})$, obtained by jointly aggregating the ratings and rankings via RR-COA (problem (20)).

As Table 5 demonstrates, there are many conflicts between the ratings-only ($\boldsymbol{x}^*$) and the rankings-only ($\boldsymbol{y}^*$) solutions; for example, paper 27 attains a top rating of 10 in the R-CA solution, whereas it is ranked as 14$^{\text{th}}$ in the R-OA solution. Such outcomes can partly be explained by the different qualities encapsulated through each input evaluation modality in the

Table 3
Evaluations of paper 18.

Judge	Attribute Ratings						Paper Ranking
	(A)	(B)	(C)	(D)	(E)	(F)	
3	3	3	4	0	2	3	4
4	5	3	4	0	5	3	4
11	6	4	4	0	5	4	4
12	7	8	8	0	7	8	2
42	2	2	2	0	3	2	4

Table 4
Evaluations of paper 26.

Judge	Attribute Ratings						Paper Ranking
	(A)	(B)	(C)	(D)	(E)	(F)	
21	8	10	8	8	5	8	3
24	8	9	8	10	7	8	1
14	7	2	3	2	2	2	5
26	8	8	7	8	8	7	3
49	10	7	6	9	9	8	1

Table 5
Aggregate evaluations for 2007 MSOM SPC.

Paper	x^*	y^*	$x^{(RR)}$	rank(x^{RR})	Paper	x^*	y^*	$x^{(RR)}$	rank(x^{RR})	Paper	x^*	y^*	$x^{(RR)}$	rank(x^{RR})
1	6.1	51	7.62	51	21	7.4	45	8.18	46	41	8.4	10	9.40	15
2	8.0	19	9.13	17	22	6.6	39	8.21	40	42	8.2	33	8.69	33
3	7.5	47	8.07	48	23	7.6	50	7.63	49	43	8.3	28	8.87	26
4	7.0	29	8.22	38	24	7.9	22	8.88	25	44	7.1	14	8.91	23
5	6.5	52	7.33	55	25	8.1	21	9.11	19	45	7.3	41	8.21	40
6	9.0	10	9.51	13	26	8.6	24	9.00	22	46	8.6	9	9.60	10
7	7.8	43	8.20	44	27	10	14	9.50	14	47	8.8	8	9.58	11
8	8.6	14	9.01	21	28	6.9	52	7.61	52	48	8.7	10	9.40	15
9	8.2	26	8.87	26	29	7.8	34	8.70	30	49	9.1	5	9.72	4
10	8.6	7	9.61	9	30	7.1	30	8.71	29	50	6.1	58	7.11	57
11	8.3	6	9.62	8	31	8.0	43	8.20	44	51	8.4	24	8.89	24
12	7.5	38	8.29	37	32	8.8	20	9.12	18	52	7.3	47	8.60	36
13	9.5	22	9.10	20	33	7.5	34	8.21	40	53	7.5	34	8.70	30
14	8.8	14	9.71	7	34	8.9	3	9.72	4	54	9.1	4	9.99	2
15	7.3	41	8.21	40	35	7.6	34	8.61	35	55	7.2	52	7.61	52
16	6.8	52	7.60	54	36	8.1	26	8.87	26	56	8.7	2	9.72	4
17	8.8	18	9.58	11	37	6.8	39	8.22	38	57	8.7	1	10.0	1
18	5.3	49	7.63	49	38	5.8	57	7.12	56	58	7.9	31	8.70	30
19	7.9	32	8.69	33	39	8.9	10	9.98	3					
20	6.0	52	6.71	58	40	7.3	46	8.08	47					

SPC. Cardinal evaluations measure average performance across all six attributes and implicitly capture judges' intensities of preference between papers, weighing equally all attributes. Conversely, ordinal evaluations capture the net preferences between papers, effectively allowing each judge to weigh and condense performance from the individual attributes differently based on what they regard as the most relevant attribute(s). Moreover, differences in assigned ranks generally do not capture intensities of preference (e.g., the preference for the first-ranked over the second-ranked paper may be marginal, but the preference for either over the third-ranked paper may be substantial). Due to the different qualities encapsulated through the two input evaluation modalities, it was not uncommon in this data set for a judge to rank a paper that performs well over all six rating attributes, but not exceptionally on any single one, lower than a paper that performs exceptionally on specific key attributes, but comparatively worse on average over all six attributes. The featured multimodal aggregation approach yields a rating-ranking pair that minimizes cumulative disagreement with the two types of input evaluations but is devoid of such conflicts.

It is important to remark that $\boldsymbol{x}^*$ and $\boldsymbol{x}^{(RR)}$ in Table 5 have higher precision than the individual attribute ratings (0.5 was the highest precision given in the attribute scores). As Section 5.1 explains, added precision is necessary to incorporate enough separation gaps in the aggregate rating. However, this does not represent a problem since R-CA and RR-COA can be solved to any rating precision, specified a priori via $\mu > 0$; herein, this parameter was set to $\mu = 0.01$.

Next, we give a specific example of objects/papers whose aggregate score in $\boldsymbol{x}^*$ and aggregate rank in $\boldsymbol{y}^*$ conflict. For instance, paper 54 has a relatively high aggregate score of 9.1, but it conflicts with others (e.g., paper 57) that have a lower aggregate score but a higher aggregate rank. Table 6 gives the evaluations received by papers 54 and 57 and their adjusted ratings, obtained by dividing the paper's rating by the respective judge's average rating. In addition, Table 7 gives the number of papers reviewed by their respective judges and the average rating these judges gave to their assigned papers. From these tables, we observe the following:

1. The ranking evaluations assigned to paper 57 seem slightly better than those assigned to paper 54.
2. The rating evaluations assigned to paper 54 were lower in magnitude than those assigned to paper 57. However, juxtaposing the paper ratings with the average rating from each respective judge suggests there was a stronger intensity of preference for paper 54 over the papers against which it was compared than for paper 57. Indeed, note that the top-3 adjusted ratings of the former are greater than those of the latter.

Table 6
Evaluations of papers 54 and 57.

Paper	Judge	Paper Rating	Paper Ranking	Adjusted Rating
54	22	7.3	1	1.47
54	25	7.0	1	1.32
54	30	6.2	1	1.25
54	32	4.6	4	0.75
57	16	7.0	1	1.04
57	17	7.4	1	1.46
57	32	7.4	1	1.21
57	57	6.3	2	1.06
57	62	6.0	1	1.17

Table 7
Statistics of judges who evaluated papers 54 and 57.

Judge	# of Papers Evaluated	AVG Rating	Judge	# of Papers Evaluated	AVG Rating
22	4	4.95	16	3	6.73
25	5	5.30	17	4	5.00
30	5	4.96	32	4	6.13
32	4	6.13	57	4	5.93
			62	4	5.15

3. The lowest rank for paper 57 was 2, while that of paper 54 was 4. Moreover, paper 54 received a below-average paper rating while paper 57 did not.

All of this suggests that paper 57 slightly edges out paper 54 when the ranking and rating evaluations are considered jointly. Indeed, in the combined aggregate rating-ranking pair, $\boldsymbol{x}^{(RR)}$ and $\text{rank}(\boldsymbol{x}^{RR})$ (the solution to RR-COA), paper 57 is rated and ranked slightly higher than paper 54; this, as discussed previously, seems appropriate. In contrast, the aggregate rating $\boldsymbol{x}^{*}$ rates paper 54 higher than 57. This analysis provides evidence that the combined rating-ranking solution (which jointly aggregates the multimodal evaluations) more effectively represents the judges' multimodal evaluations than the aggregate rating (which considers only the ratings).

Papers 14, 18, and 50 had the top-three (object-wise) contributions to the separation penalty. As noted previously and illustrated in Table 3, paper 18 elicited polarized responses: four judges gave a very low evaluation and one gave a very high evaluation. It may be prudent to further deliberate on the assigned scores/ranks in such a situation. Judges 44, 18, and 24 had the top-three (judge-wise) contributions to the separation penalty. This information suggests that these judges expressed relatively unpopular opinions. For instance, Table 8 shows that judge 44 assigned a near-perfect rating of 9.7 to paper 42 and a relatively low rating of 5.3 to paper 14 (second-worst on the judge's list), even though the solutions to R-CA, R-OA, and RR-COA all rated and/or ranked paper 42 significantly worse than paper 14. Additionally, judge 44's ratings and rankings of papers 45 and 56, whose respective evaluations are shown in Table 9, appear to be at odds with the assessments of all other judges who reviewed them and, consequently, are also at odds with the aggregate evaluations.

A potential promising line of inquiry is to examine how insights like those in the preceding two paragraphs could be used to determine the appropriateness of the initial paper-to-judge evaluation assignment and/or the existence of conflicts of interest or careless/manipulative judges. It would also be interesting to determine when the outputs from these analyses should lead to further deliberations on divergent evaluations and what specific processes can be employed. However, while these questions are relevant, they are outside the scope of this paper and are left for future work.

6.2. Generation of Synthetic Instances

Synthetic instances consist of joint ranking and rating evaluations with varying degrees of collective similarity. Individual input rankings are sampled from an adaptation of the Mallows ϕ-distribution of ranking data [47]. The standard ϕ-distribution is parametrized by a reference (i.e., ground truth) complete strict ranking $\underline{\boldsymbol{b}}$ and dispersion $\phi \in (0, 1]$, which quantify the probability of observing a complete strict ranking $\boldsymbol{b}$ as

$$P(\boldsymbol{b}) = P(\boldsymbol{b}|\underline{\boldsymbol{b}}, \phi) = \frac{1}{Z}\phi^{d_\tau(\boldsymbol{b},\underline{\boldsymbol{b}})}, \tag{31}$$

where $d_\tau(\cdot,\cdot)$ signifies the Kendall-τ [28] distance (equivalent to d_{KS} when $\boldsymbol{b}, \underline{\boldsymbol{b}}$ are complete strict rankings) and $Z = \Sigma_{\boldsymbol{b}'}\phi^{d_\tau(\boldsymbol{b}',\underline{\boldsymbol{b}})} = (1) \times (1+\phi) \times (1+\phi+\phi^2) \times \ldots \times (1+\ldots+\phi^{n-1})$ is a normalization constant. Setting $\phi = 1$ yields the (discrete) uniform distribution over the space of complete strict rankings and setting it nearer to 0 centers the distribution mass closer to $\underline{\boldsymbol{b}}$. In other words, ϕ can be said to control the proximity of $\boldsymbol{b}$ to $\underline{\boldsymbol{b}}$ and the collective similarity of multiple rankings

Table 8
Evaluations of judge 44.

Paper	Paper Rating	Paper Ranking
14	5.3	4
38	4.2	5
42	9.7	1
45	8.3	2
56	8.3	3

Table 9
Evaluations of paper 45 and 56.

Paper	Judge	Paper Rating	Paper Ranking	Paper	Judge	Paper Rating	Paper Ranking
45	23	5.2	2	56	5	6.8	1
	33	4.4	3		37	8.0	1
	40	6.2	2		44	8.3	3
	43	5.0	2		51	7.8	1
	44	8.3	2				

within a sample. It can also be said to control the difficulty of the generated instances since computation times tend to increase with ϕ. The featured experiments use these synthetic instances to assess the aggregation methodology's ability to recover an aggregate ranking that is close to the underlying ground truth $\underline{\boldsymbol{b}}$ as collective similarity weakens.

We sample instances of complete and incomplete strict rankings based on the repeated insertion model introduced by [48]. Since this sampling approach is not readily applicable for incomplete rankings, we utilize an extension developed in Yoo et al. [27] to sample from smaller projected spaces. Specifically, assuming the object set to be ranked by the k^{th} judge ($V^k_{\boldsymbol{b}}$) has been predetermined, $\boldsymbol{b}^k$ is generated according to the ϕ-distribution parametrized by $(\underline{\boldsymbol{b}}|_{V^k_{\boldsymbol{b}}}, \phi_{\boldsymbol{b}^k})$, with $b^k_i = \bullet$ for all $i \in V \setminus V^k_{\boldsymbol{b}}$—that is, $\underline{\boldsymbol{b}}|_{V^k_{\boldsymbol{b}}}$ and $\boldsymbol{b}^k|_{V^k_{\boldsymbol{b}}}$ are complete strict rankings in the projected space, and the latter of these rankings is extended to the full set of objects by assigning null values to the unranked objects ($V \setminus V^k_{\boldsymbol{b}}$). The ground truth ranking vector $\underline{\boldsymbol{b}}$ is fixed to $(1, 2, \ldots, n)$ in all the generated instances. Accordingly, the projected ground truth used to generate incomplete ranking $\boldsymbol{b}^k$ is given by $\underline{\boldsymbol{b}}|_{V^k_{\boldsymbol{b}}} = (1, 2, \ldots, |V^k_{\boldsymbol{b}}|)$.

Individual input ratings are generated using reference rating vectors and a rating error parameter. The rating scale $[L, U]$ of all inputs is $[1.0, 10.0]$. Assuming that the object set to be rated by the k^{th} judge ($V^k_{\boldsymbol{a}}$) has been predetermined, the reference rating vector $\boldsymbol{a}^k|_{V^k_{\boldsymbol{a}}}$ is set proportional to the ground truth ranking vector $\underline{\boldsymbol{b}}$ based on the number of objects rated ($|V^k_{\boldsymbol{a}}|$) and on an assigned response style. Motivated by the 2007 MSOM SPC characteristics, four response styles are defined: expanded, condensed, optimistic, and pessimistic. The first two styles differ in the expansiveness of their ranges, but each contains a balanced number of high and low *rating markers* (i.e., reference rating values); the last two styles share the same range magnitude, but each contains an unbalanced number of high or low rating markers. Table 10 lists the reference rating vectors defined for sizes $|V^k_{\boldsymbol{a}}| = 4, 5, 6, 7, 8$ for each of the four response styles. Reference rating vectors of size $|V^k_{\boldsymbol{a}}| \geq 8$ are set by assigning rating markers from $|V^k_{\boldsymbol{a}}| = 7$ to multiple objects. Objects 1 to $\lfloor \frac{|V^k_{\boldsymbol{a}}|}{7} \rfloor$ are set to the first rating marker from the respective column under $|V^k_{\boldsymbol{a}}| = 7$, objects $\lfloor \frac{|V^k_{\boldsymbol{a}}|}{7} \rfloor$ to $\lfloor \frac{2|V^k_{\boldsymbol{a}}|}{7} \rfloor$ are set to the second marker, etc. The error parameter ϵ is used to introduce random deviations from the reference rating markers and is defined as

Table 10
Setting of the (projected) ground truth rating $\boldsymbol{a}^k|_{V^k_{\boldsymbol{a}}}$ for sizes $|V^k_{\boldsymbol{a}}| = 4, 5, 6, 7, 8$ and four response styles: Expanded (**E**), Condensed (**C**), Optimistic (**O**), Pessimistic (**P**).

Rank	$\lvert V^k_{\boldsymbol{a}}\rvert = 4$				$\lvert V^k_{\boldsymbol{a}}\rvert = 5$				$\lvert V^k_{\boldsymbol{a}}\rvert = 6$				$\lvert V^k_{\boldsymbol{a}}\rvert = 7$				$\lvert V^k_{\boldsymbol{a}}\rvert = 8$			
	E	**C**	**O**	**P**	**E**	**C**	**O**	**P**	**E**	**C**	**O**	**P**	**E**	**C**	**O**	**P**	**E**	**C**	**O**	**P**
1	9.0	7.5	9.5	7.5	9.0	7.5	9.5	7.5	9.0	7.5	9.5	7.5	9.0	7.5	9.5	7.5	9.0	7.5	9.5	7.5
2	7.0	6.5	7.5	5.5	7.0	6.5	7.5	5.5	8.0	7.0	8.5	6.5	8.0	7.0	8.5	6.5	8.0	7.0	8.5	6.5
3	4.0	4.5	5.5	3.5	5.5	5.5	6.5	4.5	7.0	6.5	7.5	5.5	7.0	6.5	7.5	5.5	7.0	6.5	7.5	5.5
4	2.0	3.5	3.5	1.5	4.0	4.5	5.5	3.5	4.0	4.5	5.5	3.5	5.5	5.5	6.5	4.5	5.5	5.5	6.5	4.5
5					2.0	3.5	3.5	1.5	3.0	4.0	4.5	2.5	4.0	4.5	5.5	3.5	5.5	5.5	6.5	4.5
6									2.0	3.5	3.5	1.5	3.0	4.0	4.5	2.5	4.0	4.5	5.5	3.5
7													2.0	3.5	3.5	1.5	3.0	4.0	4.5	2.5
8																	2.0	3.5	3.5	1.5

$$\epsilon = 1.5 \ * \ \text{rand}(\{1.0, 1.5\}),$$

where rand$(\{1.0, 1.5\})$ selects one of the two scaling factors with equal probability. Given the generated ranking $\boldsymbol{b}^k$, the rating $\boldsymbol{a}^k$ is generated as follows. The k^{th} judge is first assigned one of the four responses styles. Next, the objects in $V_{\boldsymbol{a}}^k = V_{\boldsymbol{b}}^k$ are sorted based on their ascending order in $\boldsymbol{b}^k$. For each ranking position $i = 1, \ldots, |V_{\boldsymbol{b}}^k|$, an error ϵ is sampled and the object that the k^{th} judge ranks in position i receives the rating value $\underline{a}_i^k + U(-\epsilon, \epsilon)$ (i.e., the deviation term follows a continuous uniform distribution based on the sampled error parameter). The next subsection describes additional details for generating the rating and ranking aggregation instances.

Before proceeding, we discuss two practical considerations of ranking and/or rating aggregation instances and related metrics herein implemented to assess them. First, the evaluation assignments must be allocated to judges so that a direct or indirect comparison between every pair of objects in V is possible. The robustness of the object-to-judge allocation can be measured as the number of *hops* or the length in the sequence pairwise comparisons needed to obtain an implied comparison between two objects [49]. For example, for $h, i, j \in V$, if h and i are compared by one judge, i and j are compared by a different judge, and no single judge compares h and j, then there is one hop between h and i, one hop between i and j, and two hops between h and j. The robustness and reliability of the consensus aggregation solution decrease as the maximum number of hops between the object pairs increases; an allocation with a maximum hop of one over all $i, j \in V$ is ideal, and an allocation with two maximum hops is also robust. Second, it is possible for a judge's rating and ranking inputs to conflict, meaning that an object i is simultaneously ranked better and rated worse than object j (or vice versa). To quantify the degree of individual contradiction, we define the *inner distance* of judge k as the d_{KS} distance between $\boldsymbol{b}^k$ and rank$(\boldsymbol{a}^k)$ (the ranking obtained by sorting the values of $\boldsymbol{a}^k$ in non-increasing order). Note that no such contradictions can occur between the rating and ranking solution returned by RR-COA based on the enforced coupling discussed in Section 4.1. It is worth mentioning that the maximum number of hops for 2007 MSOM SPC is two, and the average inner distance is 0.06.

6.3. Analysis of Experiments on Synthetic Instances

The first experiment seeks to carry out a basic computational comparison of the Base RR-COA MILP formulation (given by (24a)-(24 h)) and the Enhanced RR-COA MILP formulation (given by (24a)-(24 h), (25a),(25b)). The formulations are tested on instances of incomplete non-strict rankings/ratings and on instances of complete non-strict rankings/ratings. This experiment primarily evaluates how the number of objects ($n = |V|$) and dispersion levels (ϕ) impact the solution times. For each defined combination of these two parameter values, 32 different instances are generated and solved by each of the two formulations within a two–hour time limit. The generated rankings do not contain ties, but the output rankings are allowed to contain ties. For simplicity, the number of ratings/rankings on each instance is set to $m = \lfloor 1.75n \rfloor$, and the expanded response style is assigned to all generated rating vectors.

For the set of incomplete non-strict ranking/rating instances, the size of each judge's evaluation subset $V_{\boldsymbol{a}}^k = V_{\boldsymbol{b}}^k$ is drawn from the discrete uniform distribution $U(4, 8)$, and the specific objects evaluated by each judge are selected randomly from universal set V. The tested numbers of objects and dispersion values are $n \in \{40, 45, 50, 55\}$ and $\phi \in \{.1, .2, \ldots, 1.0\}$, respectively. Table 11 reports the average instance and solution statistics obtained over 32 repetitions of each tested n and ϕ-value d_{KS}^{INNER}. The instance statistics are the maximum number of hops and the inner distance (label d_{KS}^{INNER}). For these instances, the maximum number of hops is exactly 2.0 and the d_{KS}^{INNER} values are between 0.073 and 0.077; both these values are thereby omitted from the table. The solution statistics are wall-clock time in seconds (label Times (s)), d_{KS} distance between the aggregate ranking and the ground truth ranking (label d_{KS}^{GT}), and the relative optimality gap percentage (label Gap%).

The same experiment is repeated on instances of all complete non-strict rankings and complete ratings as the generated inputs. Table 12 reports these results (the maximum number of hops is exactly 1.0 for all instances and the d_{KS}^{INNER} values are between 0.139 and 0.141; both these values are thereby omitted from the table). The tested numbers of objects and dispersion values for these instances are $n \in \{60, 70, 80, 90\}$ and $\phi \in \{.2, .3, \ldots, .9\}$, respectively. The narrower selection of dispersion values was selected to exclude the easiest and hardest instances; in particular, most complete RR-COA (given by (24a)-(24 h)) instances with $\phi = 1.0$ did not finish solving due to memory errors. For the tested parameter settings, the number of instances unsolved due to memory errors are reported in the last column of Table 12 (label Instances Exited); no such errors occurred for the incomplete RR-COA instances.

A few notable observations can be drawn from Tables 11 and 12. First, the ability of the models to recover the ground truth ranking diminishes as ϕ increases. Note that the values of d_{KS}^{GT} coincide when both formulations achieve optimality but may differ when either or both formulations return a suboptimal solution (due to the two–hour time limit). Second, the inner distances of incomplete RR-COA instances are roughly half the value of the complete RR-COA instances; a simple explanation for this difference is that the likelihood that an individual's evaluations are contradictory increases as more objects are evaluated. Third, computation times tend to increase with n and ϕ, and incomplete RR-COA instances tend to be more challenging to solve than complete RR-COA instances even though the latter had higher inner distance values. Fourth, the base formulation outperforms the enhanced formulation over most of the complete RR-COA instances, while the latter outperforms the former over most incomplete RR-COA instances. This is partly justified by the greater difficulty of incomplete RR-COA instances, for which incorporating the valid inequalities appears to be more worthwhile. In fact,

Table 11
Average statistics for incomplete rating and ranking instances.

n	ϕ	Times (s)		Gap%		d_{KS}^{GT}		n	Times (s)		Gap%		d_{KS}^{GT}	
		B.	**E.**	**B.**	**E.**	**B.**	**E.**		**B.**	**E.**	**B.**	**E.**	**B.**	**E.**
40	0.1	77.1	49.5	0.0	0.0	0.02	0.02	50	58.7	162.9	0.0	0.0	0.03	0.03
	0.2	68.8	62.9	0.0	0.0	0.04	0.04		139.6	330.8	0.0	0.0	0.05	0.05
	0.3	66.2	94.1	0.0	0.0	0.06	0.06		351.8	711.2	0.0	0.0	0.06	0.06
	0.4	163.8	154.2	0.0	0.0	0.10	0.10		2229.9	2052.7	0.0	0.0	0.09	0.09
	0.5	445.5	195.8	0.0	0.0	0.13	0.13		5054.7	3893.0	1.4	0.2	0.13	0.13
	0.6	2343.4	560.5	2.3	0.0	0.18	0.18		7149.0	5401.9	92.2	6.7	0.19	0.19
	0.7	4406.5	871.2	3.6	0.0	0.25	0.25		7203.9	6347.5	36.5	3.5	0.24	0.25
	0.8	5317.9	1439.8	2.6	0.0	0.32	0.32		7203.4	7029.6	29.9	6.0	0.31	0.32
	0.9	6092.9	1738.7	4.0	0.0	0.41	0.42		7203.7	7018.6	24.3	3.3	0.42	0.42
	1.0	6517.7	1984.6	4.9	0.0	0.51	0.51		7203.8	7147.5	22.8	5.7	0.50	0.49
45	0.1	52.8	73.1	0.0	0.0	0.03	0.03	55	89.0	293.3	0.0	0.0	0.03	0.03
	0.2	715.4	96.2	0.0	0.0	0.04	0.04		235.0	574.3	0.0	0.0	0.04	0.04
	0.3	110.4	254.1	0.0	0.0	0.07	0.07		1484.0	1268.6	0.0	0.0	0.07	0.07
	0.4	503.3	580.5	0.0	0.0	0.10	0.10		5128.9	3553.8	0.5	0.1	0.09	0.09
	0.5	2445.6	776.1	0.1	0.0	0.13	0.13		7047.7	6348.0	7.4	2.0	0.13	0.13
	0.6	5001.7	2105.4	31.2	0.0	0.17	0.17		7203.7	7202.8	78.7	30.1	0.19	0.19
	0.7	6962.6	4302.4	50.0	1.0	0.24	0.24		7203.8	7203.5	57.1	18.3	0.25	0.25
	0.8	7203.5	5636.3	14.9	1.2	0.32	0.33		7203.1	7201.9	37.5	12.6	0.32	0.33
	0.9	7204.2	5076.8	14.5	0.7	0.41	0.41		7203.0	7201.4	32.5	16.0	0.42	0.42
	1.0	7203.8	5572.6	15.0	0.6	0.50	0.50		7202.5	7201.3	33.6	18.3	0.51	0.51

B: Base RR-COA MILP; **E**: Enhanced RR-COA MILP

Table 12
Average statistics for complete rating and ranking instances.

n	ϕ	Times (s)		Gap%		d_{KS}^{GT}		Instances Exited	
		B.	**E.**	**B.**	**E.**	**B.**	**E.**	**B.**	**E.**
60	0.2	178.0	517.7	0.01	0.01	0.000	0.000	0	0
	0.3	208.4	386.0	0.01	0.01	0.000	0.000	0	0
	0.4	169.9	321.1	0.01	0.01	0.000	0.000	0	0
	0.5	146.6	327.4	0.01	0.01	0.000	0.000	0	0
	0.6	195.1	307.1	0.01	0.01	0.000	0.000	0	0
	0.7	260.1	427.2	0.01	0.01	0.001	0.001	0	0
	0.8	1116.1	1229.7	0.01	0.01	0.005	0.005	0	0
	0.9	6845.2	6289.0	0.06	0.04	0.020	0.020	0	0
70	0.2	710.7	1651.7	0.01	0.01	0.000	0.000	0	0
	0.3	637.1	1416.3	0.01	0.01	0.000	0.000	0	0
	0.4	615.9	1366.3	0.01	0.01	0.000	0.000	0	0
	0.5	466.2	1015.6	0.01	0.01	0.000	0.000	0	0
	0.6	504.2	915.8	0.01	0.01	0.000	0.000	0	0
	0.7	932.6	1258.5	0.01	0.01	0.001	0.001	0	0
	0.8	2112.3	2568.4	0.01	0.01	0.004	0.004	0	0
	0.9	7130.7	6510.1	0.04	0.03	0.016	0.016	7	1
80	0.2	4454.0	4200.3	0.02	0.01	0.000	0.000	0	0
	0.3	3672.9	3920.3	0.01	0.01	0.000	0.000	0	0
	0.4	2870.4	3358.8	0.01	0.01	0.000	0.000	0	0
	0.5	1807.7	2964.0	0.01	0.01	0.000	0.000	0	0
	0.6	2058.2	2187.8	0.01	0.01	0.000	0.000	1	0
	0.7	2561.9	3284.2	0.01	0.01	0.001	0.001	2	0
	0.8	4036.5	5518.0	0.01	0.01	0.003	0.003	4	0
	0.9	6990.6	7206.8	0.03	0.05	0.013	0.013	4	0
90	0.2	6589.1	7130.3	0.04	0.10	0.000	0.000	8	6
	0.3	7003.8	7142.1	0.04	0.08	0.000	0.000	6	5
	0.4	7122.3	7024.7	0.03	0.04	0.000	0.000	4	4
	0.5	6760.8	6777.7	0.02	0.02	0.000	0.000	7	5
	0.6	5382.2	5957.0	0.01	0.01	0.000	0.000	4	3
	0.7	4524.0	6147.3	0.01	0.01	0.000	0.000	3	3
	0.8	5936.8	6765.7	0.01	0.02	0.002	0.002	2	1
	0.9	7216.6	7216.3	0.05	0.12	0.011	0.011	10	8

B: Base RR-COA MILP; **E**: Enhanced RR-COA MILP

the enhanced and base formulations have comparable performances on the complete RR-COA instances with higher ϕ values; however, the enhanced formulation has a markedly superior performance for a sizable portion of the incomplete RR-COA instances.

The second experiment aims to assess the value of multimodal aggregation and to compare the computational performance of the featured optimization models. Specifically, this experiment tests the ability of RR-COA ((24a)-(24 h)), c-RR-COA ((28a)–(28j)), and R-CA (problem (9)) to recover the ground truth ranking from the generated set of noisy rating and/or ranking inputs. The R-CA consensus ranking is obtained from the non-increasing ordering of the consensus rating. The R-OA (problem (12)) model is not tested since it possesses an inherent advantage for this task. Moreover, the ability to recover the ground truth rating is not assessed since it is not well defined—it depends on the response style assigned to each judge.

The three models are tested on incomplete non-strict ranking/rating instances similar to the first experiment's. The main difference between the test instances is how the rating response styles are assigned. Two distinct rating response profiles are considered. The first profile apportioned the pessimistic response style to 55% of the judges and the expanded, condensed, and optimistic response styles equally to the remaining 45%. The second profile is similar to the first but apportions the condensed response style to 55% of the judges and the expanded, optimistic, and pessimistic response styles equally to the remaining 45%. Two other profiles with 55% expanded judges and 55% optimistic judges were tested, but they yielded similar results and are thus omitted. Another key difference from the previous experiment is that the number of judges (i.e., input evaluations) varies, specifically $m \in \{30, 40, 70, 90\}$, primarily to test the effects of different object-to-judge allocations on the ability to recover the ground truth. Other minor differences are that the number of objects is fixed to $n = 40$ and the tested dispersion values are slightly narrowed to $\phi \in \{.1, .2, \ldots, 0.8\}$, both to allow for all models to solve to optimality within two hours.

Tables 13 and 14 report the results obtained for the instances with 55% "pessimistic" judges and 55% "condensed" judges, respectively. The first general observation is that the values of d_{KS}^{GT} decrease as m increases, owing to a larger amount of information and better object-to-judge assignments resulting from the added evaluations. In fact, the values for instances with $m = 30, 40$ and $\phi = 0.1$ are approximately equal to those with $m = 70, 90$ and $\phi = 0.4$. All instances with the two highest m values achieve a maximum number of two hops, matching the object-to-judge assignment robustness of 2007 MSOM SPC—

Table 13
Average statistics for incomplete RR-COA instances with 55% "pessimistic" judges, 15% "expanded" judges, 15% "condensed" judges, and 15% "optimistic" judges.

m	ϕ	d_{KS}^{INNER}	Max Hops	Times (s)			d_{KS}^{GT}		
				RR-COA	c-RR	R-CA	RR-COA	c-RR	R-CA
30	0.1	0.062	2.78	64.83	0.11	0.05	0.063	0.065	0.116
	0.2	0.062	2.94	66.08	0.08	0.05	0.089	0.101	0.135
	0.3	0.064	2.94	77.09	0.09	0.06	0.125	0.140	0.168
	0.4	0.063	2.75	94.12	0.09	0.05	0.161	0.170	0.195
	0.5	0.061	2.88	153.37	0.09	0.05	0.218	0.220	0.234
	0.6	0.062	2.88	245.55	0.10	0.06	0.254	0.257	0.266
	0.7	0.063	2.94	473.16	0.10	0.05	0.325	0.325	0.328
	0.8	0.063	2.88	701.02	0.10	0.05	0.381	0.379	0.378
40	0.1	0.064	2.25	88.32	0.13	0.10	0.048	0.057	0.108
	0.2	0.063	2.31	70.40	0.12	0.11	0.064	0.083	0.115
	0.3	0.063	2.28	83.28	0.11	0.13	0.102	0.126	0.139
	0.4	0.063	2.38	123.87	0.12	0.13	0.132	0.159	0.168
	0.5	0.064	2.28	236.88	0.12	0.12	0.182	0.204	0.200
	0.6	0.064	2.28	687.10	0.13	0.12	0.242	0.256	0.253
	0.7	0.062	2.22	1542.45	0.13	0.12	0.284	0.305	0.299
	0.8	0.065	2.45	2872.76	0.15	0.14	0.373	0.375	0.376
70	0.1	0.064	2.00	45.66	0.16	0.19	0.025	0.043	0.077
	0.2	0.063	2.00	58.16	0.16	0.18	0.044	0.077	0.091
	0.3	0.063	2.00	68.18	0.19	0.19	0.062	0.105	0.105
	0.4	0.063	2.00	125.18	0.18	0.21	0.097	0.140	0.131
	0.5	0.066	2.00	600.00	0.19	0.19	0.134	0.173	0.153
	0.6	0.063	2.00	1587.61	0.19	0.17	0.188	0.219	0.200
	0.7	0.066	2.00	4728.64	0.22	0.19	0.236	0.266	0.247
	0.8	0.063	2.00	6079.27	0.23	0.20	0.316	0.339	0.331
90	0.1	0.063	2.00	38.58	0.24	0.23	0.017	0.040	0.066
	0.2	0.063	2.00	50.20	0.24	0.24	0.033	0.069	0.078
	0.3	0.062	2.00	63.96	0.28	0.26	0.050	0.098	0.090
	0.4	0.063	2.00	152.89	0.27	0.23	0.075	0.131	0.114
	0.5	0.063	2.00	430.42	0.31	0.26	0.117	0.166	0.141
	0.6	0.063	2.00	1740.19	0.31	0.29	0.159	0.206	0.174
	0.7	0.063	2.00	4737.39	0.25	0.28	0.214	0.255	0.227
	0.8	0.063	2.00	6524.59	0.25	0.29	0.298	0.317	0.296

Table 14
Average statistics for incomplete RR-COA instances with 55% "condensed" judges, 15% "expanded" judges, 15% "optimistic" judges, and 15% "pessimistic" judges.

m	ϕ	d_{KS}^{INNER}	Max Hops	Times (s)			d_{KS}^{GT}		
				R R - C O A	c -R R	R -C A	R R -C O A	c -R R	R -C A
30	0.1	0.083	2.91	81.33	0.12	0.06	0.071	0.073	0.130
	0.2	0.082	2.94	75.87	0.10	0.05	0.098	0.099	0.146
	0.3	0.084	2.81	80.39	0.10	0.06	0.128	0.139	0.171
	0.4	0.080	2.88	116.41	0.11	0.06	0.166	0.176	0.205
	0.5	0.081	2.78	176.23	0.11	0.05	0.216	0.220	0.238
	0.6	0.081	2.66	331.21	0.11	0.06	0.254	0.268	0.280
	0.7	0.083	2.88	334.89	0.11	0.04	0.329	0.326	0.340
	0.8	0.082	2.91	708.91	0.12	0.07	0.362	0.363	0.373
40	0.1	0.084	2.34	71.87	0.11	0.11	0.046	0.058	0.107
	0.2	0.082	2.28	72.47	0.10	0.10	0.071	0.092	0.124
	0.3	0.084	2.44	88.82	0.11	0.12	0.102	0.120	0.146
	0.4	0.083	2.28	122.27	0.11	0.11	0.141	0.168	0.184
	0.5	0.082	2.31	237.32	0.13	0.12	0.181	0.203	0.213
	0.6	0.083	2.28	449.05	0.13	0.13	0.235	0.249	0.250
	0.7	0.081	2.31	956.32	0.14	0.13	0.295	0.311	0.307
	0.8	0.083	2.28	1206.03	0.18	0.19	0.353	0.360	0.361
70	0.1	0.084	2.00	48.73	0.17	0.20	0.026	0.047	0.086
	0.2	0.083	2.00	52.47	0.18	0.20	0.043	0.074	0.101
	0.3	0.082	2.00	68.30	0.19	0.21	0.062	0.102	0.109
	0.4	0.081	2.00	132.83	0.19	0.19	0.094	0.141	0.134
	0.5	0.081	2.00	486.87	0.18	0.18	0.133	0.170	0.159
	0.6	0.082	2.00	1097.31	0.23	0.21	0.184	0.218	0.202
	0.7	0.083	2.00	3512.05	0.22	0.22	0.231	0.263	0.251
	0.8	0.082	2.00	4702.84	0.21	0.18	0.322	0.335	0.324
90	0.1	0.083	2.00	47.83	0.28	0.27	0.015	0.039	0.076
	0.2	0.081	2.00	55.39	0.31	0.28	0.031	0.071	0.086
	0.3	0.085	2.00	72.10	0.32	0.29	0.049	0.095	0.096
	0.4	0.083	2.00	134.16	0.33	0.30	0.076	0.130	0.121
	0.5	0.084	2.00	331.46	0.34	0.34	0.112	0.159	0.142
	0.6	0.083	2.00	1196.42	0.39	0.37	0.159	0.202	0.180
	0.7	0.083	2.00	4052.48	0.24	0.27	0.208	0.255	0.232
	0.8	0.083	2.00	5551.90	0.27	0.29	0.302	0.319	0.304

their respective inner distances are also similar. The maximum number of hops for instances with lower m values are between two and three on average. It is also worth remarking that computation times decreased from the two lowest to the two highest m values.

RR-COA (solved via MILP (24a)-(24 h), (25a),(25b)) is the best of the three models at recovering the ground truth ranking, particularly for instances with $\phi \leq 0.5$. Its performance over instances with higher dispersion values is less pronounced, partly because the outcomes of voting methods can become indistinguishable when there is little to no consensus in the data [8]. c-RR-COA (given by (28a)–(28j) and abbreviated as c-RR in the tables) significantly outperforms R-CA (given by (9)) in this respect, which is particularly impressive since the computational times of the two models are virtually identical. These two models solved all problems in under a second, while the RR-COA model almost reached the two–hour time limit as m and ϕ increased. Thus we conclude that c-RR-COA is an efficient and effective heuristic for solving large-scale RR-COA instances.

7. Concluding Remarks

We propose a distance-based methodology for jointly aggregating cardinal and ordinal evaluations. The methodology is designed to find a multimodal consensus evaluation—a logically coupled cardinal and ordinal evaluation pair—that minimizes the sum of the distances to the multimodal inputs. We derive linearized expressions to enforce three types of logical couplings. Furthermore, we derive a handful of optimization models to solve the rating-and-ranking aggregation variant of the methodology (which we demonstrate is NP-hard). The effectiveness of the new methodology for distributed decision-making is illustrated through a case study involving the 2007 MSOM Student Paper Competition and through synthetic instances with controllable degrees of collective similarity motivated by the case study and other practical considerations. Finally, we show compelling evidence that obtaining a combined aggregate cardinal and ordinal evaluation better represents the judges' opinions than a consensus rating that aggregates only their cardinal evaluations or a consensus ranking that aggregates only their ordinal evaluations.

The proposed methodology is founded on axiomatic distances based on social choice theory. Moreover, it is designed to adequately deal with highly incomplete evaluations and other complicating aspects of group decision-making. Aggregating

incomplete evaluations is challenging because the aggregate evaluation is especially prone to be biased by the judges' subjective scales; for example, objects assigned to a particularly strict (lenient) judge have a disadvantage (advantage) compared to objects assigned to other judges. Notably, our methodology can identify such inconsistencies in the given evaluations. This information empowers the lead decision-maker to investigate the conflicts' nature and resolve such conflicts (for example, by having the specific judges discuss and possibly resolve their assessments' inconsistencies).

The problem of aggregating complete evaluations (where all judges evaluate all objects) is a special case of the problem of aggregating incomplete evaluations (where judges may evaluate a proper subset of the objects). Therefore the methodology is also applicable to aggregating complete multimodal evaluations. Moreover, the proposed methodology may also apply to various other contexts outside of group decision-making where cardinal and ordinal evaluations over a set of entities can be obtained. For instance, [43,50] recently used modified versions of the exact multimodal aggregation model developed herein in the context of crowdsourced computation and wireless sensor networks state estimation, respectively.

Despite the proven effectiveness of the proposed model to handle multimodal inputs with different characteristics, including ties and incompleteness, it admittedly has some limitations. First, the computation time of the RR-COA model is much higher than any of the other featured models. This higher runtime is particularly prominent for instances with $\phi > 0.5$, i.e., instances with very high disagreement levels. Thus, to enhance the model's practicality, future studies will explore more sophisticated MILP solution methodologies to reduce the total computation time (e.g., further using polyhedral theory insights [45]). Second, the paper does not explicitly enforce a degree of connectivity in the pairwise-comparison graphs, i.e., maximum number of hops to indirectly compare each pair of objects. Because this aspect can significantly impact the quality and validity of the resulting aggregate evaluation, future work will seek to integrate components for robust object-to-judge allocation [49]. Third, our proposed model only interrelates ordinal and cardinal preferences; however, several other preference formats (e.g., fuzzy preference relations) could be used to express subjective evaluations. Future studies will try to address these different data modalities and create a general framework for interconnecting them within an axiomatic context.

The code used to generate the synthetic aggregation instances is available upon request.

CRediT authorship contribution statement

Adolfo R. Escobedo: Conceptualization, Methodology, Software, Formal analysis, Writing - original draft, Writing - review & editing. **Erick Moreno-Centeno:** Conceptualization, Methodology, Writing -reviewing & editing, Writing - original draft. **Romena Yasmin:** Validation, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

An unpublished manuscript titled, "Joint aggregation of cardinal and ordinal evaluations with an application to a student paper competition", available on arXiv.com, contains a preliminary version of a proper subset of the results contained in this paper.

The authors acknowledge Research Computing at Arizona State University for providing computing resources that have contributed to the research results reported within this paper. In addition, the first and third authors gratefully acknowledge funding from the National Science Foundation under grant 1850355 and from the Army Research Office under grant W911NF1910260.

References

[1] H.B. Mitchell, Data fusion: concepts and ideas, Springer Science & Business Media, 2012.
[2] D. Lahat, T. Adali, C. Jutten, Multimodal data fusion: an overview of methods, challenges, and prospects, Proceedings of the IEEE 103 (2015) 1449–1477.
[3] K. Li, X. Zhang, G. Li, A rating-ranking method for crowdsourced top-k computation, in: Proceedings of the 2018 International Conference on Management of Data, 2018, pp. 975–990.
[4] W.D. Cook, Distance-based and ad hoc consensus models in ordinal preference ranking, European Journal of Operational Research 172 (2006) 369–385.
[5] F.F. Hassanzadeh, O. Milenkovic, An axiomatic approach to constructing distances for rank comparison and aggregation, IEEE Transactions on Information Theory 60 (2014) 6417–6439.
[6] F. Brandt, V. Conitzer, U. Endriss, J. Lang, A.D. Procaccia, Handbook of computational social choice, Cambridge University Press, 2016.
[7] J. Wang, N.B. Shah, Ranking and rating rankings and ratings, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, 2020, pp. 13704–13707.
[8] H.P. Young, Condorcet's theory of voting, American Political science review 82 (1988) 1231–1244.
[9] X. Li, X. Wang, G. Xiao, A comparative study of rank aggregation methods for partial and top ranked lists in genomic applications, Briefings in bioinformatics 20 (2019) 178–189.
[10] C. Dwork, R. Kumar, M. Naor, D. Sivakumar, Rank aggregation methods for the web, in: Proceedings of the 10th international conference on the World Wide Web, New York, NY, USA, 2001, pp. 613–622.
[11] B. Fishbain, E. Moreno-Centeno, Self calibrated wireless distributed environmental sensory networks, Scientific reports 6 (2016) 24382.

[12] J.G. Kemeny, L.J. Snell, Preference ranking: An axiomatic approach, Mathematical Models in Social Science, Ginn, Boston, MA, 1962, pp. 9–23.
[13] W.D. Cook, M. Kress, Ordinal ranking with intensity of preference, Management Science 31 (1985) 26–32.
[14] K.J. Arrow, Social Choice and Individual Values, Wiley, New York, 1963.
[15] J. Bartholdi, C.A. Tovey, M.A. Trick, Voting schemes for which it can be difficult to tell who won the election, Social Choice and Welfare 6 (1989) 157–165.
[16] R. Pérez-Fernández, B. De Baets, Aggregation theory revisited, IEEE Transactions on Fuzzy Systems 29 (2020) 797–804.
[17] R. Pérez-Fernández, On an order-based multivariate median, Fuzzy Sets and Systems 414 (2021) 70–84.
[18] T. Calvo, G. Beliakov, Aggregation functions based on penalties, Fuzzy sets and Systems 161 (2010) 1420–1436.
[19] M. Gagolewski, Penalty-based aggregation of multidimensional data, Fuzzy Sets and Systems 325 (2017) 4–20.
[20] D.S. Hochbaum, A. Levin, Methodologies and algorithms for group-rankings decision, Management Science 52 (2006) 1394–1408.
[21] R.L. Keeney, A group preference axiomatization with cardinal utility, Management Science 23 (1976) 140–145.
[22] T. Saaty, A scaling method for priorities in hierarchical structures, Journal of Mathematical Psychology 15 (1977) 234–281.
[23] M. Gagolewski, R. Pérez-Fernández, B. De Baets, An inherent difficulty in the aggregation of multidimensional data, IEEE Transactions on Fuzzy Systems 28 (2019) 602–606.
[24] K.P. Bogart, Preferences structures I: Distances between transitive preference relations, Journal of Mathematical Sociology 3 (1973) 49–67.
[25] W.D. Cook, M. Kress, L.M. Seiford, An axiomatic approach to distance on partial orderings, RAIRO-Operations Research 20 (1986) 115–122.
[26] E. Moreno-Centeno, A.R. Escobedo, Axiomatic aggregation of incomplete rankings, IIE Transactions 48 (2016) 475–488.
[27] Y. Yoo, A. Escobedo, K. Skolfield, A new correlation coefficient for comparing and aggregating non-strict and incomplete rankings, European Journal of Operational Research 285 (2020) 1025–1041.
[28] M.G. Kendall, A new measure of rank correlation, Biometrika 30 (1938) 81–93.
[29] L.A. Zadeh, The concept of a linguistic variable and its application to approximate reasoning–i, Information sciences 8 (1975) 199–249.
[30] Y. Li, X. Chen, Y. Dong, F. Herrera, Linguistic group decision making: Axiomatic distance and minimum cost consensus, Information Sciences 541 (2020) 242–258.
[31] X. Chen, H. Zhang, Y. Dong, The fusion process with heterogeneous preference structures in group decision making: A survey, Information Fusion 24 (2015) 72–83.
[32] F. Chiclana, F. Herrera, E. Herrera-Viedma, Integrating three representation models in fuzzy multipurpose decision making based on fuzzy preference relations, Fuzzy sets and Systems 97 (1998) 33–48.
[33] Z. Wu, H. Liao, A consensus reaching process for large-scale group decision making with heterogeneous preference information, International Journal of Intelligent Systems 36 (9) (2021) 4560–4591.
[34] S.-M. Yu, Z.-J. Du, J.-Q. Wang, H.-Y. Luo, X.-D. Lin, Trust and behavior analysis-based fusion method for heterogeneous multiple attribute group decision-making, Computers & Industrial Engineering 152 (2021) 106992.
[35] Z.-P. Fan, J. Ma, Y.-P. Jiang, Y.-H. Sun, L. Ma, A goal programming approach to group decision making based on multiplicative preference relations and fuzzy preference relations, European Journal of Operational Research 174 (2006) 311–321.
[36] Y.-M. Wang, Z.-P. Fan, Z. Hua, A chi-square method for obtaining a priority vector from multiplicative and fuzzy preference relations, European Journal of Operational Research 182 (2007) 356–366.
[37] M. Sader, J. Verwaeren, R. Perez-Fernandez, B. De Baets, Integrating expert and novice evaluations for augmenting ordinal regression models, Information Fusion 51 (2019) 1–9.
[38] M. Tang, R. Pérez-Fernández, B. De Baets, Fusing absolute and relative information for augmenting the method of nearest neighbors for ordinal classification, Information Fusion 56 (2020) 128–140.
[39] M. Tang, R. Pérez-Fernández, B. De Baets, Distance metric learning for augmenting the method of nearest neighbors for ordinal classification with absolute and relative information, Information Fusion 65 (2021) 72–83.
[40] R.K. Ahuja, D.S. Hochbaum, J.B. Orlin, Solving the convex cost integer dual network flow problem, Management Science 49 (2003) 950–964.
[41] Y. Dong, Y. Li, Y. He, X. Chen, Preference–approval structures in group decision making: Axiomatic distance and aggregation, Decision Analysis 18 (4) (2021) 273–295.
[42] S.J. Brams, D.M. Kilgour, M.R. Sanver, A minimax procedure for electing committees, Public Choice 132 (2007) 401–420.
[43] R. Kemmer, Y. Yoo, A.R. Escobedo, R. Maciewjewski, Enhancing collective estimates by aggregating cardinal and ordinal inputs, in: Lora Aroyo, Elena Simperl (Eds.), Proceedings of the AAAI Conference on Human Computation and Crowdsourcing (HCOMP), 8, AAAI Press, 2020, pp. 73–82.
[44] Y. Yoo, A.R. Escobedo, A new binary programming formulation and social choice property for Kemeny rank aggregation, Decision Analysis 18 (4) (2021) 296–320.
[45] A.R. Escobedo, R. Yasmin, Derivations of large classes of facet-defining inequalities of the weak order polytope using ranking structures, arXiv preprint arXiv:2008.03799 (2021).
[46] A.W.K. Harzing, Response styles in cross-national survey research: A 26-country study, International Journal of, Cross Cultural Management 6 (2006) 243–266.
[47] C.L. Mallows, Non-null ranking models. i, Biometrika 44 (1957) 114–130.
[48] J.-P. Doignon, A. Pekeč, M. Regenwetter, The repeated insertion model for rankings: Missing link between two subset choice models, Psychometrika 69 (2004) 33–54.
[49] D.S. Hochbaum, A. Levin, How to allocate review tasks for robust ranking, Acta informatica 47 (2010) 325–345.
[50] J.K. Skolfield, R. Yasmin, A.R. Escobedo, L.M. Huie, A comparison of axiomatic distance-based collective intelligence methods for wireless sensor network state estimation in the presence of information injection, IEEE, 2020, pp. 1–6.