Received: 23 February 2017 | Revised: 26 November 2017 | Accepted: 27 November 2017

DOI: 10.1002/nav.21771

RESEARCH ARTICLE

WILEY

# Comparison of rank aggregation methods based on inherent ability

**Yu Xiao**[1] | **Ye Deng**[1] | **Jun Wu**[1,2] | **Hong-Zhong Deng**[1] | **Xin Lu**[1]

[1]Department of Management Science and Engineering, College of Systems Engineering, National University of Defense Technology, Changsha, Hunan 410073, P. R. China

[2]Department of Computer Science, University of California, Davis, California

**Correspondence**
Jun Wu, Department of Management Science and Engineering, College of Systems Engineering, National University of Defense Technology, Changsha, Hunan 410073, P. R. China.
Email: junwu@nudt.edu.cn

**Funding information**
National Natural Science Foundation of China, Grant/Award Numbers: 71371185, 71522014, and 71771213; Program for New Century Excellent Talents in University, Grant/Award Number: NCET-12-0141

**Abstract**

Ranking is a common task for selecting and evaluating alternatives. In the past few decades, combining rankings results from various sources into a consensus ranking has become an increasingly active research topic. In this study, we focus on the evaluation of rank aggregation methods. We first develop an experimental data generation method, which can provide ground truth ranking for alternatives based on their "inherent ability." This experimental data generation method can generate the required individual synthetic rankings with adjustable accuracy and length. We propose characterizing the effectiveness of rank aggregation methods by calculating the Kendall tau distance between the aggregated ranking and the ground truth ranking. We then compare four classical rank aggregation methods and present some useful findings on the relative performances of the four methods. The results reveal that both the accuracy and length of individual rankings have a remarkable effect on the comparison results between rank aggregation methods. Our methods and results may be helpful to both researchers and decision-makers.

**KEYWORDS**

comparison, decision analysis, experimental data generation method, inherent ability, ranking, rank aggregation

## 1 | INTRODUCTION

Motivated by applications such as world university ranking (Thieme, Prior, Tortosa-Ausina, & Gempp, 2016), web page ranking (Page, Brin, Motwani, & Winograd, 1999), and sports ranking (Coleman, 2014; Filippo, 2011), the problem of ranking has been studied extensively in the past few decades (Langville & Meyer, 2012). If there is just a single criterion for ranking, the task is relatively easy. However, in many situations, one must obtain a consensus ranking of alternatives given the individual ranking preferences of several different criteria (Dwork, Kumar, Naor, & Sivakumar, 2001; Köksalan, 1989; Lansdowne, 1996). This is known as the "rank aggregation problem" (Cook & Kress, 1990; Cook & Seiford, 1982). This problem has received significant attention recently (Köksalan, 1989; Lansdowne, 1996). Rank aggregation has penetrated many areas of decision-making and evaluation, such as meta-search engines (Dwork et al., 2001), voting systems (Obata & Ishii, 2003), and credit scoring (Bouaguel, Mufti, & Limam, 2013). Therefore, information providers and managers who rely heavily on new technology are paying significant attention to developing effective rank aggregation methods to identify the best alternatives.

Many methods for rank aggregation have been proposed over the past few decades, including the Borda's method (de Borda, 1781; Langville & Meyer, 2012), the average rank method (Langville & Meyer, 2012), the Dowdall method (Reilly, 2002), the minimum violations ranking method (Ali, Cook, & Kress, 1986; Park, 2005; Pedings, Langville, & Yamamoto, 2012; Chartier, Kreutzer, Langville, Pedings, & Yamamoto, 2010], the footrule method (Dwork et al., 2001), the Markov chain method (Dwork et al., 2001). The selection of the most appropriate aggregation method for various applications in decision-making remains a central issue in studies on rank aggregation. Thus, the evaluation and comparison of rank aggregation methods has attracted the attention of many researchers. For example, Jensen (1986)compared three continuous (ratio-scale) consensus scoring methods with the Borda-Kendall (BK) and minimum-variance (MV) ranking methods. Fields, Okudan, and Ashour (2013) developed a case study for triage prioritization to test aggregation methods. Four rank aggregation methods were applied to the prioritization data, and then an expert evaluated and judged the results in terms of practicality and acceptability.

However, most previous studies evaluated and compared rank aggregation methods using real-world datasets, which are limited not only because they are typically hard to obtain, but also because their parameters are not adjustable. Therefore, the need for a benchmark synthetic data generation method has been a recurring topic of interest in rank aggregation research. Recently, a few new synthetic data generation methods have been developed. For example, Argentini and Blanzieri (2012) developed a synthetic data generation model in which input rankings are randomly generated under a constraint to exhibit fixed values of Spearman correlation coefficients with a fixed ranking. Brancotte et al. (2015) proposed a model to create new rankings by changing the positions of elements in existing rankings. However, to the best of our knowledge, previous synthetic data generation methods can rarely generate individual rankings with adjustable length and accuracy. In practical situations, individual rankings for aggregation typically have different lengths and accuracies due to various factors. For example, in sports competitions, the tendencies of each referee are different. The lengths of different world university rankings are also nonidentical.

In this paper, we present a new experimental data generation method based on the "inherent ability" of alternatives. Our method can generate the required synthetic rankings with adjustable accuracy and length. Using the synthetic rankings and the newly proposed aggregation effectiveness criterion, we then compare four typical rank aggregation methods.

The target audience of this paper can be divided into two groups. First, for rank aggregation researchers, the synthetic rankings generation method presented here can aid them in testing if their methods are more effective than other methods. Second, for decision-makers, we provide method comparisons based on synthetic datasets, which can aid them in choosing the most appropriate rank aggregation methods for their tasks.

The remainder of this paper is organized as follows. We first introduce four typical rank aggregation methods and describe the concept of Kendall tau distance in Section 2. We then present the experimental data generation method in Section 3 and propose a new aggregation effectiveness criterion in Section 4. We compare the four rank aggregation methods in Section 5. We conclude with a summary of our contributions and a discussion of future work in Section 6.

# 2 | PRELIMINARIES

## 2.1 | Typical rank aggregation methods

Rank aggregation methods can be classified into two categories: heuristic methods and optimization methods (Argentini & Blanzieri, 2012). There are many classical methods that belong to the heuristic methods group. They aim to assign an index to each alternative that can be sorted in order to determine the consensus ranking. Alternatively, the aim of optimization methods is to find a consensus ranking that minimizes the distance to or violations with the input rankings given a particular ranking distance or violation measure, such as Kendall tau distance, Spearman footrule distance, etc. In this study, we focus on four typical aggregation methods: (1) Borda's method (BM); (2) the average rank method (ARM); (3) Dowdall method (DM); and (4) minimum violations ranking method (MVR). It should be noted that BM, ARM, and DM belong to the heuristic methods group, while MVR belongs to the optimization methods group.

### 2.1.1 | Borda's method

Borda's method is perhaps the most frequently used and simplest rank aggregation method (de Borda, 1781). For each ranking, each alternative receives a score corresponding to the position where it appears. The scores from individual rankings are then summed to create a total score, called a "Borda count," and the alternatives are sorted in descending order based on their Borda counts.

Given $k$ rankings $R_1, R_2, \ldots, R_k$, for each alternative $a \in R_i$, the alternative $a$ is first assigned a score $B_i(a) =$, which is the number of alternatives that $a$ outranks in ranking $R_i$. Next, the Borda count $B(a)$ of alternative $a$ is calculated as $\sum_{i=1}^{k} B_i(a)$. The alternatives are then sorted in descending order based on their Borda counts to create a consensus ranking.

### 2.1.2 | Average rank method

As a variant of Borda's method, the average rank method is very similar to Borda's method (Langville & Meyer, 2012). The integer positions of alternatives in rankings are averaged to calculate scores corresponding to alternatives, which are then sorted in ascending order to create a consensus ranking.

Given $k$ full rankings $R_1, R_2, \ldots, R_k$, for each alternative $a \in R_i$, alternative $a$ is first assigned a score $A_i(a) =$, which is the position of alternative $a$ in ranking $R_i$. Next, the positions of alternative $a$ in the rankings are averaged as $\frac{1}{k}\sum_{i=1}^{k} A_i(a)$. The aggregated ranking is then created by sorting the alternatives in ascending order based on their average ranks.

### 2.1.3 | Dowdall method

The Dowdall method can be considered as a "modified" form of Borda's method, but it is distinctive in some aspects. It has been widely used in political elections in many countries (Reilly, 2002). The most significant difference between Borda's method and the Dowdall method is that in Borda's method, the score assigned to each alternative varies with the size of the rankings, while for the Dowdall method, the score for an alternative is always constant and is the reciprocal of its position.

Given $k$ full rankings $R_1, R_2, \ldots, R_k$, for each alternative $a \in R_i$, alternative $a$ is first assigned a score $D_i(a) =$, which is the reciprocal of its position in ranking $R_i$. Next, the total score $D(a)$ of alternative $a$ is calculated as $\sum_{i=1}^{k} D_i(a)$. The alternatives are then sorted in descending order based on their total scores to create a consensus ranking.

### 2.1.4 | Minimum violations ranking method

The minimum violations ranking method, as its name suggests, searches specifically for consensus rankings with the minimum violations (Park, 2005). Typically, the binary integer linear program (BILP) formulation of the MVR problem is the preferred way to find the optimal consensus ranking (Chartier et al., 2010; Langville & Meyer, 2012; Pedings et al., 2012). Denote by $x_{ij}$ the decision variables that determine whether alternative $a_i$ should be ranked above alternative $a_j$. Specifically:

$$x_{ij} = \begin{cases} 1, & \text{if alternative } a_i \text{ is ranked above } a_j \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

Some constraints must be added to force the matrix $X$ to have the properties that meet the basic needs for producing a unique ranking of the $n$ alternatives:

$$x_{ij} + x_{ji} = 1; \quad x_{ij} + x_{jk} + x_{ki} \le 2 \tag{2}$$

Given $k$ rankings $R_1, R_2, \ldots, R_k$ of the $n$ alternatives, we define the following ranking scores for any pair of objects (Langville & Meyer, 2012):

$$c_{ij} = \left(\#of \text{ rankings with } a_i \text{ above } a_j\right) - \left(\#of \text{ rankings with } a_i \text{ below } a_j\right). \tag{3}$$

The objection of MVR is to find the consensus ranking maximizing the conformity among input rankings. In terms of ranking scores $c_{ij}$ and variables $x_{ij}$, this becomes:

$$\max \sum_{i=1}^{n} \sum_{j=1}^{n} c_{ij} x_{ij}. \tag{4}$$

BILPs are typically solved with a technique called branch and bound, which uses a series of linear programming (LP) relaxations of the problem to form a tree to narrow down the process of stepping through the discrete solution space (Langville & Meyer, 2012). When the branch and bound procedure terminates with an optimal solution $X^*$, we can obtain a MVR consensus ranking by sorting the column sums of $X^*$ in ascending order (Langville & Meyer, 2012). It is worth noting that if $c_{ij} = c_{ji}$ for some $(i,j)$, the alternatives $i$ and $j$ in the optimal ranking can be swapped without changing the objective value. If this is so, then an alternate optimal ranking is one that has these two alternatives swapped, which could lead to the existence of multiple MVR rankings. We can use the *Tie Detection* algorithm (Langville & Meyer, 2012) to identify all the optimal rankings and groups of *special alternatives* that can be swapped without changing the objective value. Accordingly, a unique optimal MVR ranking with ties can be obtained in the way that each group of *special alternatives* share the same rank position (Langville & Meyer, 2012; Pedings et al., 2012). As a matter of fact, this unique optimal MVR ranking with ties in some rank position can be viewed as the average of all optimal rankings (Chartier et al., 2010). For the purpose of convenience, we use this unique optimal MVR ranking to evaluate the performance of MVR in this paper.

## 2.2 | Kendall tau distance

Kendall tau distance (Kemeny, 1959) is a metric that counts the number of pairwise disagreements between two ranking lists. The distance between two rankings $R_1$ and $R_2$ is defined as:

$$K(R_1, R_2) = |\{(i,j) \mid i < j, R_1(i) < R_1(j), \text{ but } R_2(i) > R_2(j)\}|. \tag{5}$$

The larger the distance, the more dissimilar the two rankings are. $K(R_1, R_2)$ will be zero if the two rankings are identical and $n(n-1)/2$ if $R_1$ is the reverse of $R_2$ ($n$ is the size of rankings $R_1$ and $R_2$).

# 3 | EXPERIMENTAL DATA GENERATION METHOD

## 3.1 | Accuracy of individual rankings

Consider a rank aggregation problem with $N$ voters and $M$ alternatives. We assume that there exists a ground truth ranking of the alternatives. It can be the latent ranking of the actual strengths of each alternative that individual voters and by extension, the rank aggregation itself are attempting to

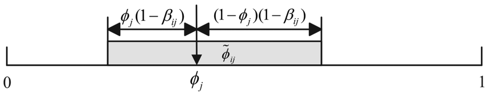

**FIGURE 1** The displayed inherent ability of $a_j$ for voter $b_i$

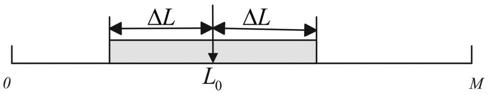

**FIGURE 2** The length of individual ranking $R_i$

estimate given the displayed abilities of those alternatives. To acquire the ground truth ranking of the alternatives, we denote the inherent ability of the alternative $a_j$ as $\phi_j$. It may be a certain attribute of $a_j$, such as the height of a person, the quality of a product, or the English level of a student. We assume that the inherent ability $\phi_j$ follows a uniform distribution in the region $[0,1]$. The ground truth ranking of alternatives based on their inherent abilities is denoted $R_0 = [r_1, r_2, \ldots, r_M]$, where $r_j$ is the true position of alternative $a_j$ based on $\phi_j$. A larger inherent ability of an alternative corresponds to a higher position. Considering the fact that voters may not be perfectly aware of $\phi_j$ in practice for a variety of reasons, we denote the displayed inherent ability of alternative $a_j$ for voter $b_i$ as $\tilde{\phi}_{ij}$. We assume that voters rank alternatives based on the displayed inherent ability $\tilde{\phi}_{ij}$. The ranking of alternatives given by voter $b_i$ is denoted $R_i = [\tilde{r}_{i1}, \tilde{r}_{i2}, \ldots, \tilde{r}_{iM}]$, where $\tilde{r}_{ij}$ is the position of alternative $a_j$ ranked by voter $b_i$ based on $\tilde{\phi}_{ij}$. The task of rank aggregation is to combine all individual rankings $R_i$ into a consensus ranking $\hat{R}$.

We assume that $\tilde{\phi}_{ij}$ is a random variable following a uniform distribution in the region $[\phi_j - \phi_j(1-\beta_{ij}), \phi_j + (1-\phi_j)(1-\beta_{ij})]$, as shown in Figure 1. The parameter $\beta_{ij} \in [0,1]$ represents the accuracy of the inherent ability of alternative $a_j$ for the voter $b_i$. The larger $\beta_{ij}$ is, the narrower the distribution region is and the more accurate the displayed inherent ability $\tilde{\phi}_{ij}$ is. This corresponds to the accuracy of voter $b_i$ in ranking alternative $a_j$. There are two extreme cases. When $\beta_{ij} = 1$, we have $\tilde{\phi}_{ij} = \phi_j$. In other words, voter $b_i$ can rank the alternatives exactly according to the inherent ability of each alternative. When $\beta_{ij} = 0$, $\tilde{\phi}_{ij}$ is a random variable with a uniform distribution in the region $[0,1]$. This means that voter $b_i$ ranks the alternatives randomly. For the purpose of convenience, we assume that the displayed accuracy $\beta_{ij}$ for all alternatives and voters are identical, meaning $\beta_{ij} = \beta$ for all $i \in [1,N]$ and $j \in [1,M]$.

### 3.2 | Length of individual rankings

It should be noted that, for a variety of reasons, voters may only rank a small number of alternatives, meaning only an incomplete list of alternatives is compared and ordered into a partial ranking $R_i$. We define $\tilde{r}_{ij}$ as the position of alternative $a_j$ within the set of alternatives that are ranked by voter $b_i$. If voter $b_i$ does not rank alternative $a_j$, we define $\tilde{r}_{ij} = 0$. The length of the ranking $R_i$ is denoted $L_i = |\left\{\tilde{r}_{ij} | \tilde{r}_{ij} > 0, 1 \le j \le M\right\}|$. It follows that $0 \le L_i \le M$. We assume that $L_i$ is a random variable following a uniform distribution in the region $[L_0 - \Delta L, L_0 + \Delta L]$, as shown in Figure 2. The parameter $L_0 \in [0,M]$ represents the baseline length of the individual rankings and $0 \le \Delta L < L_0$ represents the variation in the individual ranking lengths. If $\Delta L = 0$, all rankings have the same length.

## 4 | MEASURE OF RANK AGGREGATION EFFECTIVENESS

A simple but effective measure of rank aggregation effectiveness is crucial for evaluating rank aggregation methods. The traditional measure of rank aggregation effectiveness is defined as the sum of the Kendall tau distances $S$ between the aggregated ranking $\hat{R}$ and all individual rankings $R_i$. This can be written as follows:

$$S = \sum_{i=1}^{N} K(\hat{R}, R_i), \tag{6}$$

where $K$ is the Kendall tau distance. The smaller the value of $S$ is, the more effective the rank aggregation method is. The essence of the traditional measure of rank aggregation effectiveness is to characterize the centrality of the aggregated ranking with respect to the individual rankings.

Here, we propose measuring the effectiveness of rank aggregation methods by using the Kendall tau distance $D$ between the aggregated ranking $\hat{R}$ and the ground truth ranking $R_0$. This can be written as follows:

$$D = K(\hat{R}, R_0). \tag{7}$$

The smaller the value of $D$ is, the more effective the rank aggregation method is. Rather than the centrality of the aggregated ranking, our proposed measure $D$ characterizes the correctness of the aggregated ranking.

## 5 | COMPARISON OF TYPICAL RANK AGGREGATION METHODS

We now evaluate the four typical rank aggregation methods introduced in Section 2. We first generate various individual synthetic rankings using our experimental data generation method, and then aggregate individual rankings using the four

**TABLE 1** The effectiveness measure $D$ versus the ranking accuracy $\beta$ with various $L_0$, where $N = 1000$, $M = 100$, and $\Delta L = 0.3L_0$. The values in bold and italic text represent the best effectiveness measures among the four methods. The results are averaged over 100 independent trials

| | $L_0$=10 | | | | $L_0$=30 | | | | $L_0$=50 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\beta$ | BM | ARM | DM | MVR | BM | ARM | DM | MVR | BM | ARM | DM | MVR |
| 0.10 | 1666 | ***1238*** | 1473 | 1270 | 1157 | 776 | ***696*** | 794 | 867 | 621 | ***462*** | 636 |
| 0.30 | 721 | ***436*** | 625 | 465 | 400 | ***235*** | 267 | 259 | 279 | ***178*** | 180 | 200 |
| 0.50 | 403 | ***202*** | 371 | 237 | 210 | ***102*** | 166 | 133 | 138 | ***75*** | 110 | 103 |
| 0.70 | 295 | 116 | 307 | ***113*** | 149 | ***50*** | 143 | 75 | 97 | ***35*** | 95 | 53 |
| 0.90 | 254 | 79 | 277 | ***58*** | 123 | 26 | 130 | ***24*** | 77 | 16 | 83 | ***13*** |

rank aggregation methods. We then compare the effectiveness of the four rank aggregation methods. We focus specifically on the impact of ranking accuracy and ranking length on the comparison results. All experiments are repeated 100 times in order to obtain stable results. We implement a host of analysis and statistical dispersion tools and observe the low relative standard deviations (RSD) for various methods and parameters, which should guarantee that the experimental data is reliable. For example, with $N = 1000$, $M = 100$, $\beta = 0.5$, $L_0 = 40$, and $\Delta L = 12$, we obtain $RSD = 0.0994$ for BM, $RSD = 0.11184$ for ARM, $RSD = 0.10634$ for DM, and $RSD = 0.0958$ for MVR.

To test the significance of the differences between the four methods, we implement a host of nonparametric Friedman tests for multiple paired samples from the experimental results using SPSS. The results suggest that there are significant differences between the four methods. For example, the Asymp. Sig. ($P$-value) for the data in Table 1 is 0.000, which is much smaller than 0.05. We obtain the same conclusion for the effectiveness measure $S$.

## 5.1 | Impact of ranking accuracy

To investigate the impact of ranking accuracy $\beta$ on the evaluation of rank aggregation methods, we implement numerical experiments and present the measures of effectiveness of rank aggregation methods $D$ with various $\beta$ and $L_0$ in Figure 3. The specific data is also presented in Table 1, where we emphasize the best effectiveness measure among the four methods using bold and italic text. We find that, with increasing values in ranking accuracy $\beta$, the values of the effectiveness measure D decrease rapidly, meaning the effectiveness of rank aggregation methods is better when ranking accuracy $\beta$ is higher, which agrees with our intuition.

Most importantly, we find that there is a threshold for ranking accuracy $\beta$. When ranking accuracy rises above the threshold, ARM and MVR perform much better than the other two methods. This suggests that the accuracy of individual rankings has a remarkable impact on the comparison results between rank aggregation methods. It is worth noting that the threshold depends on the baseline length of the individual rankings $L_0$ and the variation in individual ranking lengths $\Delta L$. For example, the threshold is 0.3 with $L_0 = 30$ and $\Delta L = 9$, while the threshold is 0.5 with $L_0 = 50$ and $\Delta L = 15$.

## 5.2 | Impact of ranking length

To investigate the impact of the baseline length of the individual rankings $L_0$ on the evaluation of rank aggregation

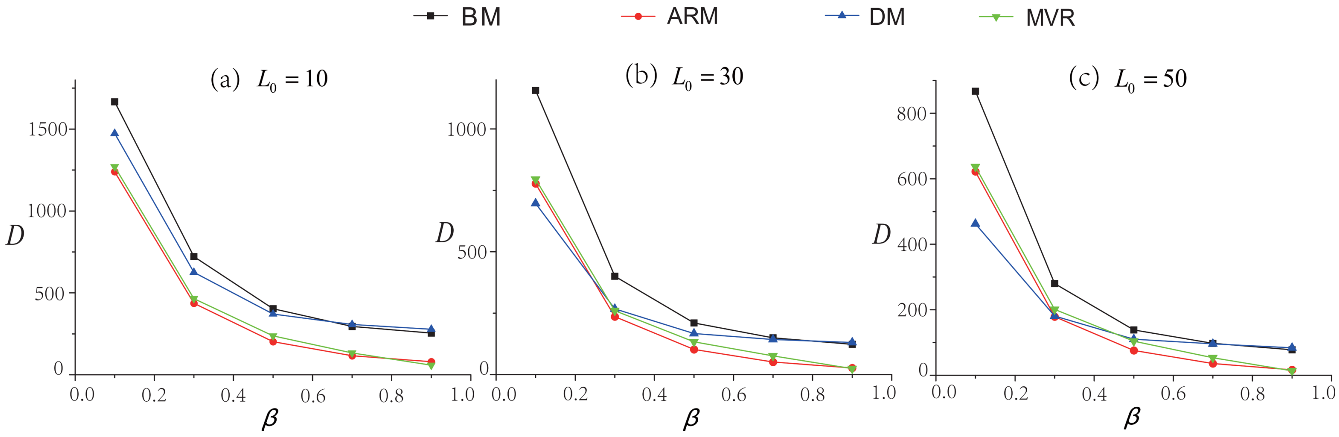

**FIGURE 3** The effectiveness measure $D$ versus the ranking accuracy $\beta$ with various $L_0$, where $N = 1000$, $M = 100$, and $\Delta L = 0.3L_0$. The results are averaged over 100 independent trials [Color figure can be viewed at wileyonlinelibrary.com]

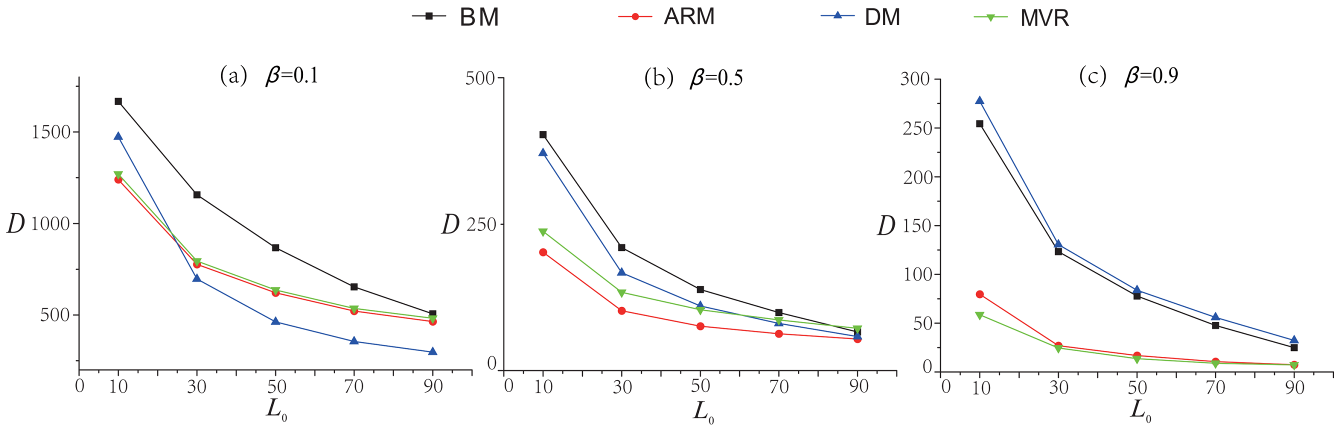

**FIGURE 4** The effectiveness measure $D$ versus the baseline length $L_0$ with various $\beta$, where $N = 1000$, $M = 100$, and $\Delta L = 0.3L_0$. The results are averaged over 100 independent trials [Color figure can be viewed at wileyonlinelibrary.com]

methods, we implement numerical experiments and present the measures of effectiveness of rank aggregation methods $D$ with various $\beta$ and $L_0$ in Figure 4. The specific data is also presented in Table 2. We observe that, with increasing values of the baseline length of rankings $L_0$, the values of the effectiveness measure $D$ decrease rapidly. It is unsurprising and intuitive that the effectiveness of rank aggregation methods is better with larger baseline lengths of the individual rankings $L_0$. This is because the larger baseline lengths of the rankings $L_0$ correspond to more complete evaluation information. Furthermore, we find an interesting phenomenon where, in the case of low ranking accuracy ($\beta = 0.1$), when the baseline length of the individual rankings is large enough ($L_0 > 30$), DM becomes the preferred method among the four methods.

To investigate the impact of the variation in the individual ranking lengths on the evaluation of rank aggregation methods, we present the effectiveness measure $D$ as a function of the variation in individual ranking lengths $\Delta L$ in Figure 5. The specific data is also presented in Table 3. We find that the variation in the individual ranking lengths has a remarkable effect on the effectiveness of ARM and DM. An increase in $\Delta L$ leads to the decrease in the effectiveness of ARM and DM. However, BM and MVR are both very robust against increases in $\Delta L$. In the case of high ranking accuracy ($\beta = 0.9$), when the variation in individual ranking lengths is large enough ($L_0 > 30$), MVR will overtake ARM and become the preferred method among the four methods. Our results suggest that the variation in the individual ranking lengths $\Delta L$ should also be taken into consideration when evaluating rank aggregation methods.

## 5.3 | Comparison of two effectiveness measures

We now explore the differences between the traditional measure of rank aggregation effectiveness $S$ and the proposed measure $D$. We first use the two measures to evaluate the four typical rank aggregation methods: BM, ARM, MVR, and DM. We then rank the four methods based on the two measures. A smaller rank represents better effectiveness.

To characterize the differences between $S$ and $D$, we calculate the Kendall tau distances between the ranking of the four methods based on $S$ and $D$, which are shown as colored lumps in Figure 6. The color of each lump corresponds to the value of the Kendall tau distance. A blue lump corresponds to a small value for Kendall tau distance, meaning that the two measures are similar, while a red lump

**TABLE 2** The effectiveness measure $D$ versus the baseline length $L_0$ with various $\beta$, where $N = 1000$, $M = 100$, and $\Delta L = 0.3L_0$. The values in bold and italic text represent the best effectiveness measures among the four methods. The results are averaged over 100 independent trials

| | $\beta$=0.1 | | | | $\beta$=0.5 | | | | $\beta$=0.9 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $L_0$ | BM | ARM | DM | MVR | BM | ARM | DM | MVR | BM | ARM | DM | MVR |
| 10 | 1666 | ***1238*** | 1473 | 1270 | 403 | ***202*** | 371 | 237 | 254 | 79 | 277 | ***58*** |
| 30 | 1157 | 776 | ***696*** | 794 | 210 | ***102*** | 166 | 133 | 123 | 26 | 130 | ***24*** |
| 50 | 867 | 621 | ***462*** | 636 | 138 | ***75*** | 110 | 103 | 77 | 16 | 83 | ***13*** |
| 70 | 653 | 521 | ***355*** | 535 | 99 | ***62*** | 80 | 86 | 47 | 10 | 55 | ***8*** |
| 90 | 506 | 463 | ***296*** | 482 | 66 | ***53*** | 58 | 72 | 24 | ***7*** | 32 | ***7*** |

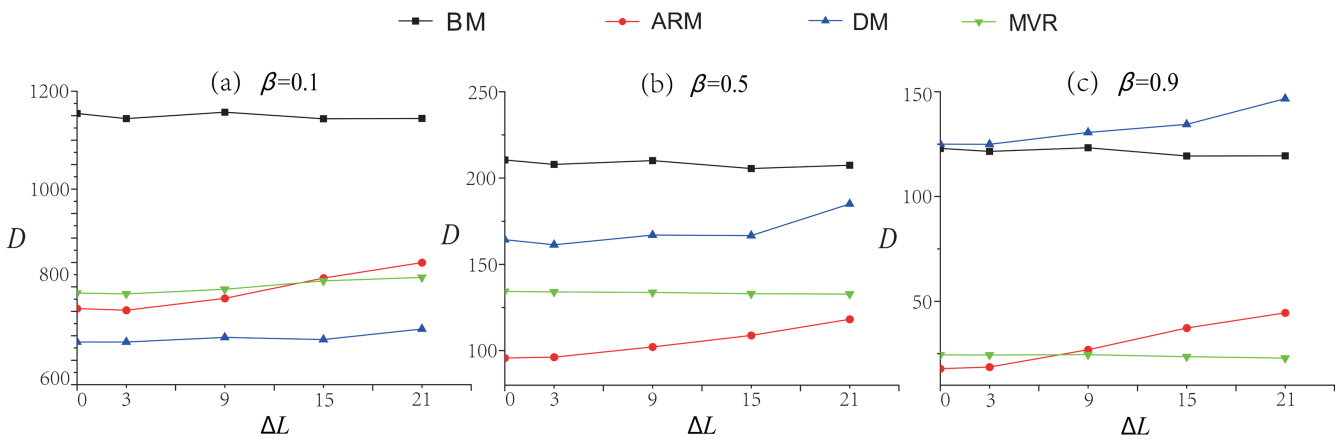

**FIGURE 5** The effectiveness measure $D$ versus the variation in ranking length $\Delta L$ with various $\beta$, where $N = 1000$, $M = 100$, and $L_0 = 30$. The results are averaged over 100 independent trials [Color figure can be viewed at wileyonlinelibrary.com]

corresponds to a large value for Kendall tau distance, meaning that the two measures are significantly different. It is clear that there are significant differences between the traditional measure of rank aggregation effectiveness $S$ and the proposed measure $D$, particularly in cases of low ranking accuracy $\beta$.

In order to present the mismatches between the ranks of the four methods based on $S$ and $D$ intuitively, we present the ranking pairs of the four methods with various parameter combinations in Figure 7. The left side is the rank based on $S$ and the right side is the rank based on $D$. A line represents a pair of ranks based on the measures $S$ and $D$. For example, with $N = 1000$, $M = 100$, $L_0 = 30$, and $\Delta L = 9$, we see that the rank of DM based on $D$ is 4 and the rank of DM based on $S$ is 3. Thus, there is a line between rank “4” on the left side and rank “3” on the right side. The thickness of each line is proportional to the number of occurrences of the rank pair with various parameter combinations. Horizontal lines indicate that there is no difference between the ranks based on the two measures, while oblique lines represent a mismatch of ranks based on the two measures. As shown in Figure 7, numerous oblique lines indicate a significant difference between the proposed measure $D$ and the traditional measure $S$ for rank aggregation effectiveness.

## 6 | CONCLUSIONS AND DISCUSSION

Studies on rank aggregation have received increasing attention in the past few decades. In this study, we focused on the evaluation of various rank aggregation methods. We accomplished this evaluation by developing an experimental data generation method. In this method, we introduced the concept of the inherent ability of alternatives, from which we can obtain the ground truth ranking of the alternatives and generate the required individual synthetic rankings with adjustable accuracy and length. We have proposed a new measure for the effectiveness of rank aggregation methods by calculating the Kendall tau distance between the aggregated ranking and the ground truth ranking. We have demonstrated that both the accuracy and length of individual rankings have a significant effect on the comparison results between rank aggregation methods and emphasized that there is a significant difference between the proposed effectiveness measure and the traditional effectiveness measure.

Using the experimental data generation method, we compared four typical rank aggregation methods: the Borda’s method (BM), the average rank method (ARM), the Dowdall method (DM), and the minimum violations ranking (MVR).

**TABLE 3** The effectiveness measure $D$ versus the variation in ranking length $\Delta L$ with various $\beta$, where $N = 1000$, $M = 100$, and $L_0 = 30$. The values in bold and italic text represent the best effectiveness measures among the four methods. The results are averaged over 100 independent trials

| | $\beta$=0.1 | | | | $\beta$=0.5 | | | | $\beta$=0.9 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\Delta L$ | BM | ARM | DM | MVR | BM | ARM | DM | MVR | BM | ARM | DM | MVR |
| 0 | 1154 | 755 | ***687*** | 787 | 210 | ***95*** | 164 | 134 | 122 | ***17*** | 125 | 24 |
| 3 | 1143 | 755 | ***687*** | 785 | 207 | ***96*** | 161 | 134 | 121 | ***18*** | 124 | 24 |
| 9 | 1157 | 776 | ***696*** | 794 | 210 | ***102*** | 166 | 133 | 123 | 26 | 130 | ***24*** |
| 15 | 1143 | 818 | ***692*** | 812 | 205 | ***108*** | 166 | 132 | 119 | 37 | 134 | ***23*** |
| 21 | 1144 | 849 | ***713*** | 819 | 207 | ***118*** | 185 | 132 | 119 | 44 | 146 | ***22*** |

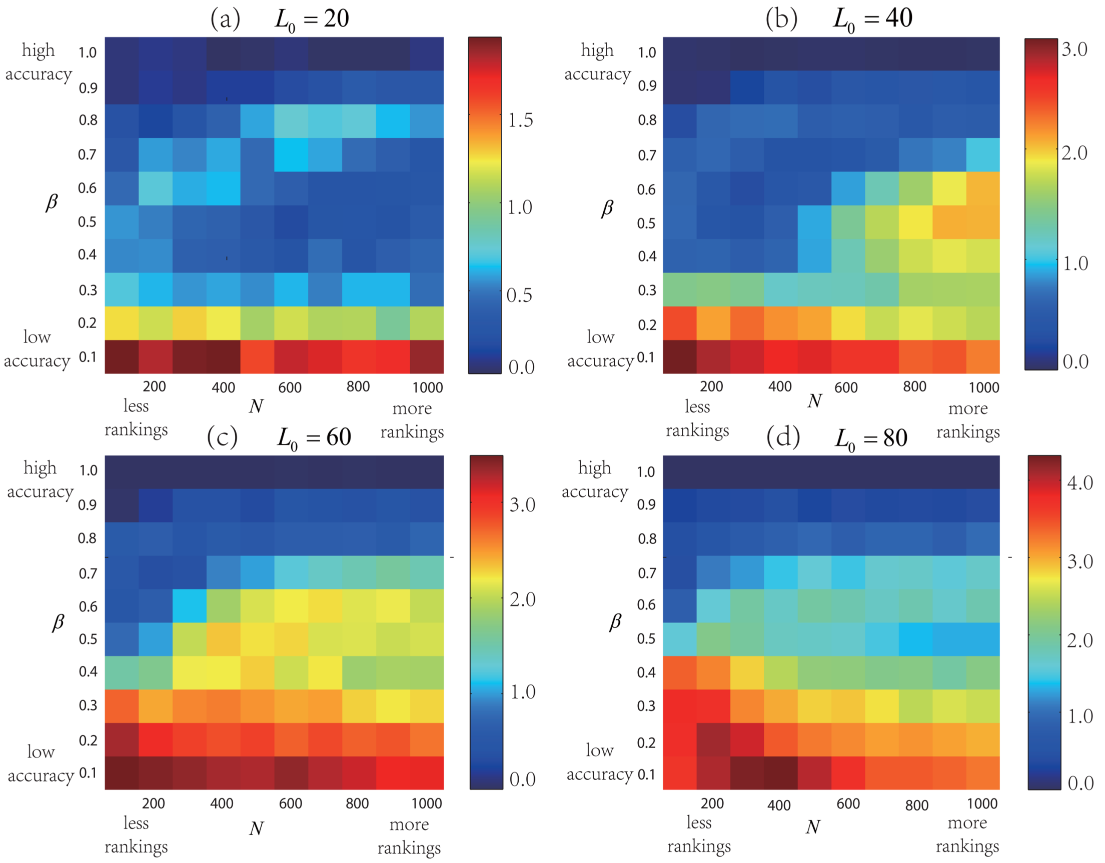

**FIGURE 6** The Kendall tau distances between the rankings of four methods derived from $S$ and $D$ with various $N$, $\beta$, and $L_0$, where $M = 100$ and $\Delta L = 0.5L$. The results are averaged over 100 independent trials [Color figure can be viewed at wileyonlinelibrary.com]

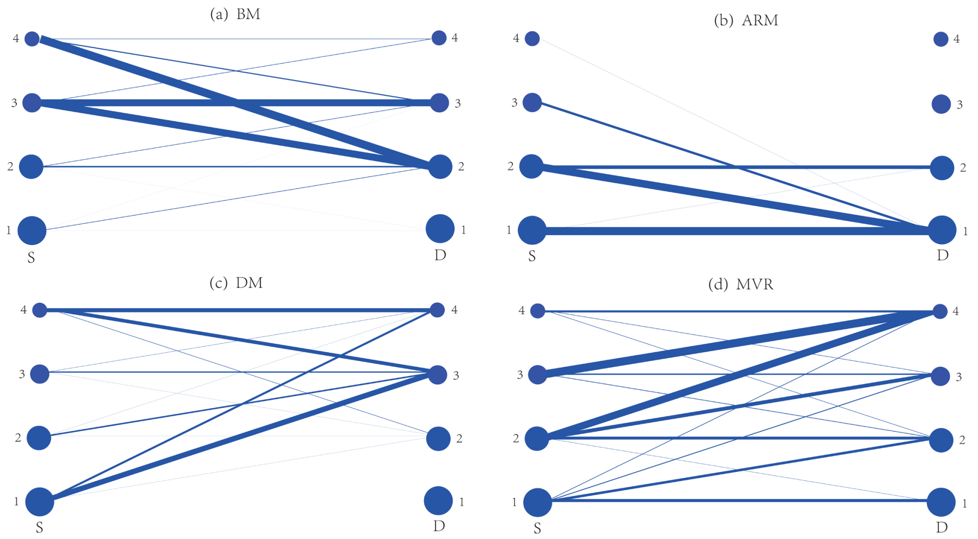

**FIGURE 7** The rank pairs of the four methods with various parameter combinations. The left side is the rank based on $S$ and the right side is the rank based on $D$. A line represents a pair of ranks based on the two measures $S$ and $D$. The thickness of each line is proportional to the number of occurrences of the rank pair with various parameter combinations [Color figure can be viewed at wileyonlinelibrary.com]

We have obtained some useful findings regarding the relative performance of the four methods: (1) in the case of low ranking accuracy and large baseline length of the individual rankings, DM is the preferred method; (2) in the case of high ranking accuracy, ARM and MVR are much better than BM and DM; (3) in the case of high ranking accuracy and large variation in the individual ranking lengths, MVR is the preferred method; (4) MVR is quite reliable in most cases. Although we have not provided a systematic and comprehensive comparison of all the rank aggregation methods, the main contribution of this paper is the presentation of a general theoretical framework for the comparison of rank aggregation methods, which can be widely adapted according to the task requirements.

Our results provide both theoretical insights into the effectiveness of rank aggregation methods and practical knowledge regarding method selection for users. First, we developed and tested a synthetic ranking generation method that can aid rank aggregation researchers in testing whether or not their methods are effective. Second, our findings provide a number of new and interesting directions for future research, such as exploring new parameters that may impact the effectiveness of rank aggregation. Finally, the conclusions we have made may be helpful in many areas for managing information from various sources, such as the areas of meta-search engines and voting systems. A particularly meaningful conclusion from our research is that an effective aggregated ranking reflects the ground truth ranking of alternatives. This concept may be pursued in rank aggregation research in the future.

## ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China (grant numbers 71371185, 71522014, and 71771213) and the Program for New Century Excellent Talents in University (grant number NCET-12-0141). We thank Suo-Yi Tan and Yang Zhang for their helpful insights.

## ORCID

*Jun Wu* http://orcid.org/0000-0002-1154-6071

## REFERENCES

Ali, I., Cook, W. D., & Kress, M. (1986). On the minimum violations ranking of a tournament. *Management Science*, *32*(6), 660–672.

Argentini, A., & Blanzieri, E. (2012). Ranking Aggregation Based on Belief Function. In: S. Greco, B. Bouchon-Meunier, G. Coletti, M. Fedrizzi, B. Matarazzo, R. R. Yager (Eds), Advances in Computational Intelligence. IPMU 2012. Communications in Computer and Information Science, Springer, vol. 299, pp. 511–520.

Bouaguel, W., Mufti, G. B., & Limam, M. (2013). Rank aggregation for filter feature selection in credit scoring. In: Prasath R., Kathirvalavakumar T. (Eds), Mining Intelligence and Knowledge Exploration. Lecture Notes in Computer Science, Springer, vol. 8284, pp. 7–15).

Brancotte, B., Yang, B., Blin, G., Cohen-Boulakia, S., Denise, A., & Hamel, S. (2015). Rank aggregation with ties: Experiments and analysis. *Proceedings of the VLDB Endowment*, *8*(11), 1202–1213.

Chartier, T. P., Kreutzer, E., Langville, A. N., Pedings, K., & Yamamoto, Y. (2010). Minimum violations sports ranking using evolutionary optimization and binary integer linear program approaches. In A. Bedford & M. Ovens (Eds.), *Proceedings of the Tenth Australian Conference on Mathematics and Computers in Sport* (pp. 13–20). New South Wales, Australia: MathSport (ANZIAM).

Coleman, B. J. (2014). Minimum violations and predictive meta-rankings for college football. *Naval Research Logistics (Nrl)*, *61*(1), 17–33.

Cook, W. D., & Kress, M. (1990). A data envelopment model for aggregating preference rankings. *Management Science*, *36*(11), 1302–1310.

Cook, W. D., & Seiford, L. M. (1982). On the borda-kendall consensus method for priority ranking problems. *Management Science*, *28*(6), 621–637.

de Borda, J. C. (1781). *Mémoire sur les élections au scrutin*. *Histoire De L'Académie Royale Des Sciences*. Paris.

Dwork, C., Kumar, R., Naor, M., & Sivakumar, D. (2001). Rank aggregation methods for the web. In *WWW 2001, Proceedings of the 10th international conference on World Wide Web, Hong Kong* (pp. 613–622). New York, NY: ACM.

Fields, G. E., Okudan, E. B., & Ashour, O. M. (2013). Rank aggregation methods comparison: A case for triage prioritization. *Expert Systems with Applications*, *40*(4), 1305–1311.

Filippo, R. (2011). Who is the best player ever? a complex network analysis of the history of professional tennis. *PLoS ONE*, *6*(2), e17249.

Jensen, R. E. (1986). Comparison of consensus methods for priority ranking problems. *Decision Sciences*, *17*(2), 195–211.

Kemeny, J. G. (1959). Mathematics without numbers. *Daedalus*, *88*(4), 577–591.

Köksalan, M. M. (1989). Identifying and ranking a most preferred subset of alternatives in the presence of multiple criteria. *Naval Research Logistics*, *36*(4), 359–372.

Langville, A. N., & Meyer, C. D. (2012). *Who is #1? The science of rating and ranking*, Princeton, NJ: Princeton University Press.

Lansdowne, Z. F. (1996). Ordinal ranking methods for multicriterion decision making. *Naval Research Logistics*, *43*(5), 613–627.

Obata, T., & Ishii, H. (2003). A method for discriminating efficient candidates with ranked voting data. *European Journal of Operational Research*, *151*(1), 233–237.

Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). *The PageRank Citation Ranking: Bringing Order to the Web* (Technical Report 1999-0120). Computer Science Department, Stanford University.

Park, J. (2005). On minimum violations ranking in paired comparisons, arXiv preprint physics/0510242.

Pedings, K. E., Langville, A. N., & Yamamoto, Y. (2012). A minimum violations ranking method. *Optimization and Engineering*, *13*(2), 349–370.

Reilly, B. (2002). Social choice in the south seas: Electoral innovation and the borda count in the pacific island countries. *International Political Science Review*, *23*(4), 355–372.

Thieme, C., Prior, D., Tortosa-Ausina, E., & Gempp, R. (2016). Value added, educational accountability approaches and their effects on schools' rankings: Evidence from Chile. *European Journal of Operational Research*, *253*(2), 456–471.

**How to cite this article:** Xiao Y, Deng Y, Wu J, Deng H-Z, Lu X. Comparison of rank aggregation methods based on inherent ability. *Naval Research Logistics*. 2017;00:1–10. https://doi.org/10.1002/nav.21771