2021 ■ Vol. 17 ■ no. 2

Using the weighted Kendall Distance to analyze rank data in psychology

Johnny van Doorn [a], Holly A. Westfall [b] & Michael D. Lee [b]

[a]Department of Psychological Methods, University of Amsterdam
[b]Department of Cognitive Sciences, University of California, Irvine

Abstract ■ Although the Kendall distance is a standard metric in computer science, it is less widely used in psychology. We demonstrate the usefulness of the Kendall distance for analyzing psychological data that take the form of ranks, lists, or orders of items. We focus on weighted extensions of the metric that allow for heterogeneity of item importance, item position, and item similarity, as well showing how the metric can accommodate missingness in the form of top-k lists. To demonstrate how the Kendall distance can help address research questions in psychology, we present four applications to previous data. These applications involve the recall of events on September 11, people's preference rankings for the months of the year, people's free recall of animal names in a clinical setting, and expert predictions involving American football outcomes.

Keywords ■ Ordinal data, Modeling tool, Rank correlation.

JohnnyDoorn@gmail.com

10.20982/tqmp.17.2.p154

Acting Editor ■ Roland Pfister (Universität Würzburg)

Reviewers
■ Two anonymous reviewers

Introduction

Rank data are found in most areas of psychological science. Any task that involves sequences of behavior, such as recalling items from memory or solving a problem through a series of decisions and actions, yields rank order data (e.g., Hamada, Nakayama, & Saiki, 2020; Healey & Kahana, 2014). Other common examples range from psychophysics (Gordon, 1924) to consumer choice preferences (Adomavicius, Bockstedt, & Curley, 2015; Milosavljevic, Navalpakkam, Koch, & Rangel, 2012).

An often-used statistical tool to analyze rank data is a rank correlation, such as Kendall's τ (Kendall, 1938) or Spearman's ρ (Spearman, 1904). The goal of these methods is to quantify the strength of a monotonic relation between two variables, without assuming this relation to be linear. The rank correlation coefficient is then frequently used to test hypotheses related to the presence or absence of such a relation. However, such a procedure often overlooks the wealth of information embedded in the value of the rank correlation coefficient. In computer science, for instance, rank correlations are a popular metric for aggregating search engine results, fighting spam, and word association (Beg & Ahmad, 2003). Whereas psychological science predominantly uses the rank *correlation* to test for an association between two variables, the field of computer science focuses on the (non-standardized) rank *distance* to quantify degrees of similarity between two or more observed sequences of data points. In doing so, the distance metric becomes a function of the data that can, in turn, be used for further quantitative analysis.

In this article, we aim to bridge the gap between developments in computer science and psychological science by underscoring the Kendall distance metric as a useful tool for analyzing psychological data.[1] First, we outline the basic distance metric, which has sometimes been used in psychology (e.g., Lee, Steyvers, & Miller, 2014; Brandt, Conitzer, Endriss, Lang, & Procaccia, 2016; Selker, Lee, & Iyer, 2017). Second, we discuss three extensions introduced by Kumar and Vassilvitskii (2010) that enable the weighting of item importance, item position, and item similarity, which are rarely used in psychology. Third, we illustrate how the Kendall distance can be modified to accommodate missingness in the data in the form of top-k lists, as introduced by Fagin, Kumar, and Sivakumar (2003). Each extension is first illustrated using a toy example, and then

[1]We focus on the Kendall distance as a modeling tool, rather than a hypothesis testing framework.

demonstrated more fully in practical applications to existing data sets in psychology previously collected to address specific research questions. In order to increase the ease of application of the discussed algorithms, we include a plug and play R-script, available at https://osf.io/6k9t8/. The R-script is demonstrated for each toy example application.

The Kendall Distance

Introduced by Kendall (1938), the Kendall distance metric, often written as τ, is a popular rank-based coefficient for comparing two vectors of data points. It is based on the number of adjacent pairwise swaps required to transform one vector into the other.

In order to present the notation[2] and computation of the Kendall distance and its extensions, we use a small toy example where two people are asked to rank $n = 4$ sodas—Coke, Pepsi, Sprite, and Fanta—in terms of tastiness. Let the ranking of person A be $A = (Coke, Pepsi, Fanta, Sprite)$, and the ranking of person B be $B = (Pepsi, Coke, Sprite, Fanta)$. We denote the ith item of A with A_i, such that $A_i = Coke$ when $i = 1$. Next, we denote the ranking of item c with $\sigma_A(c)$ for person A, and $\sigma_B(c)$ for person B. For instance, $\sigma_A(c) = 1$ and $\sigma_B(c) = 2$ for $c = Coke$. Combining these two notations allows us to denote the rank of the ith item in A, for person B. For instance, $\sigma_B(A_i) = 2$ when $i = 1$, because the first item in A (i.e., Coke) is ranked second by person B.

With these definitions in hand, we can compute the Kendall distance between person A and B. In order to sort B in such a way that it is identical to A, we need to swap $Coke$ and $Pepsi$, and then $Fanta$ and $Sprite$. In this example, the Kendall distance is therefore equal to 2. As a consequence, the Kendall distance is often referred to as the bubble sort distance (Shaw & Trimble, 1963).[3] Table 1 provides an illustration of this sorting procedure.

In order to obtain the correlation coefficient, the distance is then standardized to be in the interval $[-1, 1]$, however, we focus on the distance in this article. The minimum value for the distance is 0, indicating perfect correspondence, and the maximum value for the distance is equal to $n(n-1)/2$, where n is the length of A and B.

Another way of calculating the Kendall distance is by comparing the ranks of items A_i and A_j in the vector B, for $i < j$. If item $\sigma_B(A_i)$ is greater than $\sigma_B(A_j)$, this means that person B ranked items A_i and A_j in the reverse order compared to person A. We refer to this as an *inversion*. A formal definition is given by the formula:

$$\tau = \sum_{1 \leq i < j \leq n} \left[\sigma_B(A_i) > \sigma_B(A_j)\right], \tag{1}$$

which counts the number of pairwise inversions.

We can compute the Kendall distance using the `calcTopTau` function from https://osf.io/6k9t8/. The function takes in two vectors of ranked objects, and computes the Kendall distance. If the input is not numeric, the function automatically assigns numeric values to the ranked objects. To improve readability, the function is demonstrated here with string input. The conversion to numeric values is arbitrary, so we caution users to be aware of the mapping from string to numeric values when applying the function.

```
calcTopTau(
   x = c("Coke", "Pepsi", "Fanta", "Sprite"),
   y = c("Pepsi", "Coke", "Sprite", "Fanta"))

## [1] 2
```

We now discuss four extensions of the Kendall distance that have the potential to be especially useful for analyzing psychological data. The first three extensions were first introduced by Kumar and Vassilvitskii (2010), and the fourth extension by Fagin et al. (2003).

Item Weights

As presented by Kumar and Vassilvitskii (2010) Item-specific weights may be incorporated in the distance metric. In the basic definition, the cost of swapping two items is set to 1, such that swapping two items adds 1 to the metric. However, it could be the case that some items contribute more, or less, to the dissimilarity between the soda preferences of person A and B. For instance, we could theorize that disagreement in taste is more important in the ranking of Fanta than for other sodas. For instance, the marketing team of Fanta may want two people who rank Fanta differently to be recognized as being more dissimilar than two people who rank Coke differently. In such cases, we can use the item specific weights w, where w_i denotes the cost of performing a swap that contains item A_i. This enables us to model different items as contributing more, or less, to differences between the rankings represented by the vectors A and B.

Formally, the extension to include item importance is given by the formula:

$$\tau = \sum_{1 \leq i < j \leq n} w_{A_i} w_{A_j} \left[\sigma_B(A_i) > \sigma_B(A_j)\right]. \tag{2}$$

In order to include item weights in the `calcTopTau` function, a numeric vector can be supplied to the `itemWeights` argument. The order of the weights should correspond to the alphabetical order of the ranked

[2] We follow the notation of Kumar and Vassilvitskii (2010).

[3] See also https://www.youtube.com/watch?v=lyZQPjUT5B4 and https://www.geeksforgeeks.org/bubble-sort/ for accessible introductions.

Table 1 ■ The two vectors A and B, and the adjacent pairwise swaps needed to transform B into A: B^1 denotes B after one swap, and B^2 denotes B after two swaps. Therefore, the Kendall distance between A and B equals 2.

A	B	B^1	B^2
Coke	Pepsi	Coke	Coke
Pepsi	Coke	Pepsi	Pepsi
Fanta	Sprite	Sprite	Fanta
Sprite	Fanta	Fanta	Sprite

items, if these are not numeric. Below is an example of weighting Fanta as twice as important as the other items:

```
calcTopTau(
   x = c("Coke", "Pepsi", "Fanta", "Sprite"),
   y = c("Pepsi", "Coke", "Sprite", "Fanta"),
   itemWeights = c(1, 2, 1, 1))
```

```
## [1] 3
```

Position Weights

Another extension focuses on weighting different positions in a ranking, rather than different items. This can be achieved by making the cost of performing a swap dependent on the position i on which an inversion occurs (Kumar & Vassilvitskii, 2010). We can imagine a situation in which one's favorite soda is more important in determining taste preference than one's least favorite soda: if an inversion occurs early in B, this should lead to a greater value of the Kendall distance than if an inversion occurs at the end of B.

In order to assign these weights, we first define p:

$$p_i = p_{i-1} + \delta_i,$$

where $p_1 = 1$. The weight δ_i denotes the cost of a pairwise swap of an item in the ith position. It therefore represents the importance of that position, relative to the first position. This weight can either be assigned arbitrarily, or through a specific algorithm. One popular method is called discounted cumulative gain (DCG; Järvelin & Kekäläinen, 2002), where the weights are calculated as a logarithmic function of the item positions:

$$\delta_i = \frac{1}{\log(i+1)} - \frac{1}{\log(i+2)}.$$

The intuition behind the DCG weighting is that the item at position i is about twice as important in determining the dissimilarity between A and B than the item at position $i-1$, so that when the item is sorted, its swaps have a lower cost. For an illustration of this, see Table 2.

With δ and p defined, we can now calculate the average cost of moving item i in B to the position of that item in A, remembering that this can involve multiple pairwise swaps. This average cost is:

$$\bar{p_i} = \frac{p_i - p_{\sigma_B(A_i)}}{i - \sigma_B(A_i)}.$$

For instance, the cost of moving item Coke from position 2 to position 1 in B is calculated as

$$\bar{p_1} = \frac{p_1 - p_{\sigma_B(Coke)}}{1 - \sigma_{B(Coke)}} = \frac{p_1 - p_2}{1-2} = \frac{1 - 1.189}{1-2} = 0.189.$$

The general incorporation of position weights is provided by the formula:

$$\tau = \sum_{1 \le i < j \le n} \bar{p_i}\bar{p_j} \left[\sigma_B(A_i) > \sigma_B(A_j)\right]. \tag{3}$$

The `calcTopTau` function includes several possibilities for weighting according to item position using the `posWeights` argument. First, the discounted cumulative gain, or its reverse, can be used. Since there are two swaps in the example – one using the first two items, and one using the last two items – the DCG and reverse DCG give identical results. Second, the position weights can be specified manually. These options can also be used in combination with the `nTOPK` argument, which can be used to only consider the first n observations.

```
# Using the Discounted Cumulative Gain
calcTopTau(
   x = c("Coke", "Pepsi", "Fanta", "Sprite"),
   y = c("Pepsi", "Coke", "Sprite", "Fanta"),
   posWeights = "DCG")
```

```
## [1] 0.03967739
```

```
# Using the reverse Discounted Cumulative Gain
calcTopTau(
   x = c("Coke", "Pepsi", "Fanta", "Sprite"),
   y = c("Pepsi", "Coke", "Sprite", "Fanta"),
   posWeights = "revDCG")
```

```
## [1] 0.03967739
```

Table 2 ■ Values of the position weights δ and p_i for the soda example, with δ calculated using the DCG algorithm.

i	A	B	δ	p
1	Coke	Pepsi	-	1
2	Pepsi	Coke	0.189	1.189
3	Fanta	Sprite	0.1	1.289
4	Sprite	Fanta	0.063	1.352

```
# Manual specification of position weights
calcTopTau(
   x = c("Coke", "Pepsi", "Fanta", "Sprite"),
   y = c("Pepsi", "Coke", "Sprite", "Fanta"),
   posWeights = 4:1)

## [1] 10

# Only consider the first two observations
calcTopTau(
   x = c("Coke", "Pepsi", "Fanta", "Sprite"),
   y = c("Pepsi", "Coke", "Sprite", "Fanta"),
   nTOPK = 2)

## [1] 1
```

Similarity Weights

The third extension of the Kendall distance takes into account the similarities and differences between items. This means that, when two items are considered highly similar, the cost of swapping these two items is lower than the cost of swapping two items that are considered to be more different from one another. In our sodas example, this can be used to model the high similarity between Coke and Pepsi.[4] As such, the inversion of Coke and Pepsi has a lower cost than the inversion of Fanta and Sprite.

In order to incorporate item similarities, we define the distance matrix D, where element D_{ij} determines how similar items A_i and A_j are. When this is set to 0, items A_i and A_j are identical; as the values are set to the large values, the item pairs become more different.

The distance matrix is incorporated as follows in the formula for the Kendall distance:

$$\tau = \sum_{1 \leq i < j \leq n}^{n} D_{ij} \left[\sigma_B(A_i) > \sigma_B(A_j) \right]. \tag{4}$$

In order to include a distance matrix in the `calcTopTau` function, a numeric matrix can be supplied to the `distMat` argument. The order of the weights should correspond to the alphabetical order of the ranked items, if these are not numeric.[5] Below is an example of incorporating the similarity between Coke and Pepsi:

```
distMat <- matrix(1, ncol = 4, nrow = 4)

distMat[1, 3] <- distMat[3, 1] <- 0.1
calcTopTau(
   x = c("Coke", "Pepsi", "Fanta", "Sprite"),
   y = c("Pepsi", "Coke", "Sprite", "Fanta"),
   distMat = distMat)

## [1] 2
```

Top-k Lists

Lastly, we discuss comparing top-k lists. When comparing two lists of k items, it may be the case that not all items appear on both lists. It could be that there is no predetermined set of items to rank. For example, instead of asking person A and B to rank four sodas, they could have been asked to list their top 4 favorite sodas. It could also be the case that one ranking contains missing information. For example, even if the same four sodas are being ranked, one person might only list their top three. This sets the current extension apart from the previous three extensions: whereas the other extensions are modeling choices, the top k extension is driven by the nature of the data.

Suppose that we observe the responses $A = (Coke, Pepsi, Fanta, Sprite)$ and $B = (7up, Sprite, GingerAle, Pepsi)$. In such a case, we cannot determine for all items if an inversion has occurred due to some items only appearing in one of the lists. A method introduced by Fagin et al. (2003) can be used to model the missingness of items.

The approach identifies four cases of how two items may appear in A and B, and outlines the cost of a swap z. We present these four cases for the toy example:

1. Both items appear in A and B (e.g., Sprite and Pepsi). Since person A prefers Pepsi and person B prefers Sprite, this is the traditional case of an inversion and therefore $z = 1$.
2. Both items appear in A, but only one item appears in B (e.g., Pepsi and Fanta). Since person B only includes Pepsi, we can conclude that they prefer Pepsi over Fanta. If person A shares this preference, $z = 0$. Otherwise, this is an inversion and therefore $z = 1$.

[4] The authors acknowledge that some readers might wildly disagree with this statement.
[5] The code available on https://osf.io/4ej6s/ provides a method of specifying the similarity weights by name instead of index.

3. One item appears only in A, and the other item appears only in B (e.g., Coke and 7up). In a similar reasoning to the previous case, we know that person A prefers Coke over 7up and person B prefers 7up over Coke, because at least those sodas appear in the list. This is an inversion and we therefore set $z = 1$.
4. Both items appear in A, but neither appear in B (e.g., Coke and Fanta). Here there is no information on whether person B prefers Coke or Fanta, since neither appear in B. As a first option, Fagin et al. (2003) outline the optimistic approach, which is setting $z = 0$. In other words, this gives person B the "benefit of the doubt," and assumes that if they had included Coke and Fanta, they would have expressed the same preference as person A. Alternatively, the pessimistic approach sets $z = 1$, and assumes person B would have expressed the opposite ordering of the items. As such, z can be conceptualized as the probability of person B preferring Coke over Fanta. For instance, specifying $z = \frac{1}{2}$ corresponds to a neutral approach, in which there is an equal probability for person B expressing the same or reverse order for the items. This still takes into account the missingness, while not making a statement about how the items would be ranked if they would be included in B.

Adding this extension to the Kendall distance formula gives:

$$\tau = \sum_{1 \leq i < j \leq n} z_{ij} \left[\sigma_B(A_i) > \sigma_B(A_j) \right], \tag{5}$$

where z_{ij} depends on the specific pair of items (i, j), and its value determined as outlined above.

The missingness parameter can be specified in the `calcTopTau` function by setting the `missingProb` argument:

```
calcTopTau(
   x = c("Coke", "Pepsi", "Fanta", "Sprite"),
   y = c("7up", "Sprite", "Ginger Ale", "Pepsi"),
   missingProb = 0) # optimistic

## [1] 11

calcTopTau(
   x = c("Coke", "Pepsi", "Fanta", "Sprite"),
   y = c("7up", "Sprite", "Ginger Ale", "Pepsi"),
   missingProb = 1) # pessimistic

## [1] 13
```

All of the extensions presented above can be combined to form the weighted partial Kendall distance:

$$\tau = \sum_{1 \leq i < j \leq n} w_i w_j \bar{p}_i \bar{p}_j D_{ij} z_{ij} \left[\sigma_B(A_i) > \sigma_B(A_j) \right]. \tag{6}$$

Applications

We have now defined the full metric that is capable of modeling item importance, item position, and item similarity, while also accommodating missingness in top-k lists. In this section, we present a series of four applications of the Kendall distance to previous psychological data, demonstrating how the various extensions can improve data analysis to address the motivating research questions.

Item Weights: Recall of Events on September 11

In order to study memory reconstruction, Altmann (2003) considered six events that occurred on September 11, 2001. The events, in their true temporal order, were (1) One plane hits the World Trade Center, (2) A second plane hits the World Trade Center, (3) One Plane crashes into the Pentagon, (4) One tower at the World Trade Center collapses, (5) One Plane crashes in Pennsylvania, and (6) A second tower at the World Trade Center collapses.

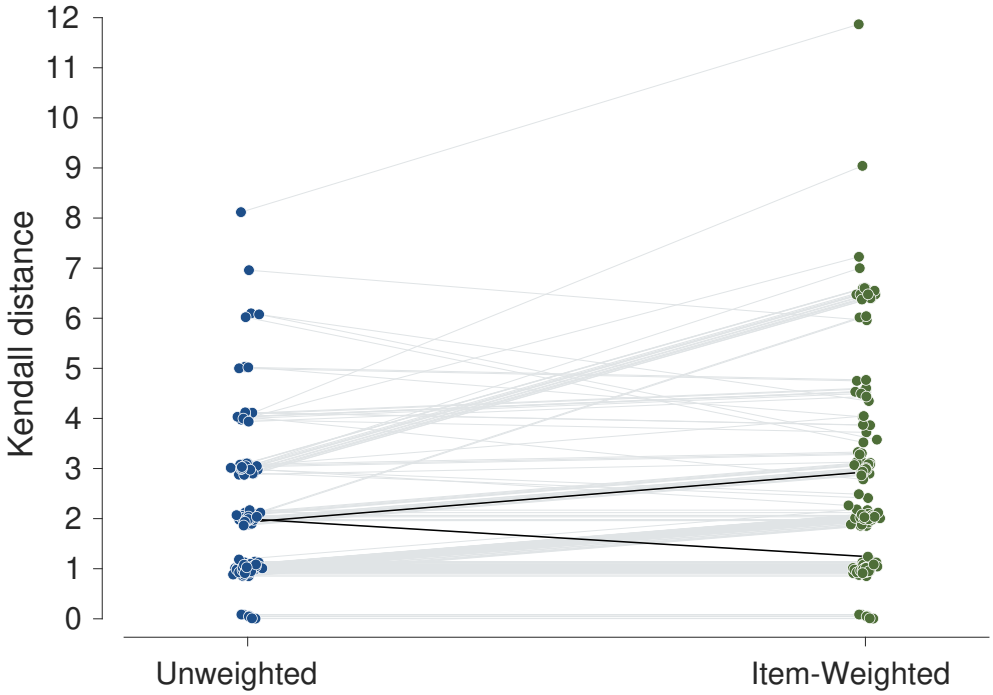
The participant responses consist of individual's recalled temporal orderings of these events. The Kendall distance provides a natural single measure of response accuracy for each participant. However, as noted by Altmann (2003), the correct ranking of some of these events need not be driven by memory, but can be determined by logic. For example, it can be deduced that the planes hitting the World Trade Center occurs before the tower collapsing, and that the first plane hits before the second plane. In contrast, correctly recalling when the plane crash in Pennsylvania occurred needs to be memory driven. Thus, when a participant incorrectly orders the two planes hitting the towers, this can be due to poor memory or poor reasoning, while incorrectly ranking the Pennsylvania crash is more likely due to poor memory.

These considerations mean that if the research goal is to study memory ability in recall, rather than logic reasoning skill, events (3) and (5) should be weighted more heavily than items (1), (2), and (4). For example, if we consider the responses from two specific participants in the Altmann (2003) data, who recalled orders: (A) Plane 2, Pentagon, Plane 1, Tower 1, Pennsylvania, Tower 2 (B) Plane 1, Pentagon, Plane 2, Pennsylvania, Tower 1, Tower 2

Both of these participants have the same number of inversions relative to the ground truth, and therefore yield an identical unweighted Kendall distance of 2. However, participant A makes logical errors while participant B does not. As a consequence, assigning a weight of 2 to the memory driven items, and a weight of $1/2$ to the logic driven items, changes the accuracy measures to 1.25 for participant A and 3 for participant B.

Figure 1 shows the change in the Kendall distance resulting from including item weights for 158 participants

Figure 1 ■ Unweighted and item-weighted Kendall distance for 158 participants from the Altmann (2003) study of memory for the order of events on September 11. The distances are between the participants' responses and the ground truth. Each point in the unweighted column and item-weighted column corresponds to a participant, jittered around the Kendall distance measure. The same participant for each measure is connected by a gray line. Participants A and B are highlighted by black lines. As a result of the weighting, Participant A sees a decrease of the Kendall distance, whereas Participant B sees an increase of the Kendall distance.



from Altmann (2003). The standard unweighted measure is shown on the left, and the item-weighted measure is shown on the right, with lines connecting the same participant under each measure. It is clear that the recall accuracy of participants can increase, decrease, or stay the same once item weights are incorporated. It is also clear that the use of item weights also gives the Kendall distance greater resolution as a measure of accuracy. Without weighting, there theoretically are 15 possible outcomes for the Kendall distance, 9 of which are observed in the Altmann (2003) data. With item weighting there are 61 possible outcomes, including fractional counts, 21 of which are observed.

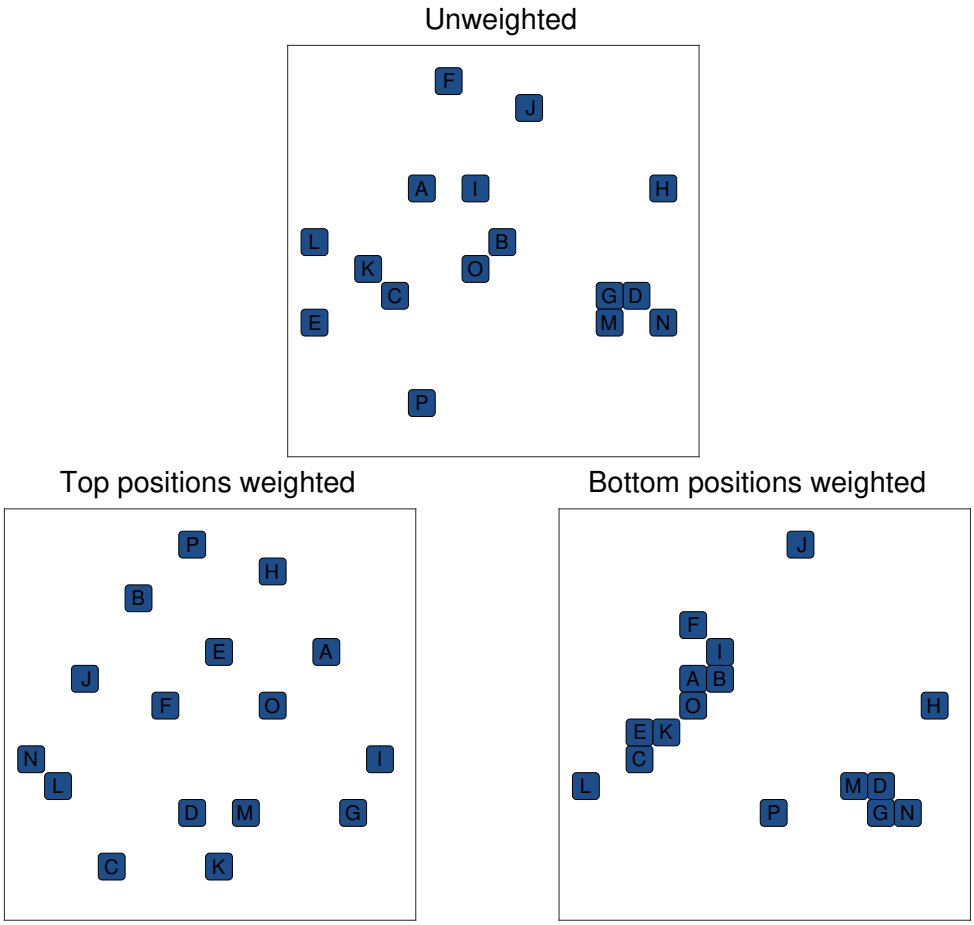
Position Weights: Month Preference

In the previous example, the participant responses were compared to a true ranking, in order to determine their accuracy. However, participants' responses can also be compared to each other, in order to determine similar response patterns. Accordingly, our second application involves people's preferences for the months of the year, as collected by the crowd-source opinion web site ranker.com. A total of 16 people ranked the 12 months from best worst.

A natural research question addressed by these data is whether there are individual differences in people's preference patterns. For instance, some people prefer the winter months to the summer months, while others may prefer summer to winter. One exploratory approach to identifying such patterns is through data visualization. We apply the multidimensional scaling MDS:, Borg and Groenen (1997) algorithm to the pairwise Kendallâ€™s distances between people, using spaces of just two dimensions. This allows for a simple visualization that may reveal clusters of people based on the similarity of their preferences (i.e., groups of participants whose Kendall distance scores are small with respect to each other).

There are two extensions of the Kendall distance that are potentially useful here. First, we can model the adjacent months as being fairly similar to each other. We can therefore reduce the cost of swapping, for instance, January and February from 1 to 0.5. Secondly, we can use position weights to capture assumptions about whether peo-

Figure 2 ■ MDS visualization based on the unweighted, top-position weighted, and bottom-positions weighted Kendall distances between people's preferences for the months of the year. Each square/letter indicates a different participant. The distances between the squares/letters represent their similarities according to the Kendall distance.



ple's most or least favorite months are more indicative of their preference. For example, consider the rankings provided two ranker.com users: (A) Dec, Jun, Oct, May, Jul, Nov, Aug, Apr, Sep, Mar, Feb, Jan (B) May, Oct, Nov, Dec, Sep, Jun, Jul, Apr, Aug, Mar, Feb, Jan Their favorite months are rather different, but their least favorite months are very similar. Whether these people are regarded as having similar preferences depends on the weighting given to their favorite months, as compared to their least favorite months.

Figure 2 presents the MDS visualizations for all of the ranker.com people, considering three scenarios. The top panel shows the MDS visualization for the preference rankings that are weighted by similarity but are unweighted by position. The lower-left panel shows the MDS visualization for the preference rankings where the DCG algorithm was used to weight the best months more heavily. The bottom-right panel shows the MDS visualization for the preference rankings where the reverse DCG algorithm is applied, in order to weigh the worst months more heavily.

In this way, the difference between the bottom-left and bottom-right visualizations is based on whether the most favored or least favored months are treated as the most important in determining the similarity between people's preferences. Accordingly, in terms of the specific examples presented earlier, person A and person B are further apart in the top and bottom-left panels of Figure 2 than they are when the weighting is changed to emphasize the least favorite months, as in the bottom-right panel.

It is striking that the MDS visualization based on weighting the least favorite months, shown in the bottom-right panel, reveals a clear cluster structure. There is a di-

vide between people who dislike the cold winter months, in the left half of the plot, and people who dislike the hot summer months, in the right half of the plot. The other visualizations lack this clear cluster structure, suggesting that focusing on the months people like the least is a good way to understand the group structure of their preferences.

Similarity Weights: The Free Recall of Animals

Our third application involves measuring performance on a free recall memory task in a clinical setting, and focuses on the use of similarity weights. The data were collected using the Mild Cognitive Impairment Screen (MCIS: Shankle, Mangrola, Chan, & Hara, 2009), one component of a routine assessment of Alzheimer's patients in a neurodegenerative disorders clinic. As part of this assessment, patients complete a triadic comparison task for nine animal names, where each of the animals is presented in a triad with each of the other animals and the patient must determine which of the three animal names is least like the other two. After a delay, patients complete a surprise free recall task of those nine animal names.

One important research goal is to identify and understand the different free recall response patterns. There is evidence that the semantic relationships between the animals influences the order in which their names are recalled (Bousfield & Sedgewick, 1944; Bousfield, 1953; Romney, Brewer, & Batchelder, 1993). In particular, it is common for the recalled list to be made up of sub-sequences of semantically-related animal names. For example,"zebra", "giraffe", "elephant", and "tiger" are likely to be recalled consecutively, as a cluster of African zoo animals. In clinical settings, the exact order in which a cluster like this is recalled is less important than the fact it is recalled largely as a cluster, since this suggests semantic memory is intact.

As a concrete example, consider the recall data for three people: (A) Elephant, Giraffe, Sheep, Rat, Monkey, Chimpanzee, Rabbit, Zebra, Tiger (B) Rat, Sheep, Giraffe, Zebra, Elephant, Monkey, Chimpanzee, Tiger, Rabbit (C) Rat, Chimpanzee, Zebra, Giraffe, Elephant, Tiger, Rabbit, Sheep, Monkey

The unweighted Kendall distance between A and B is 11, between A and C is 18, and between B and C is 13, which implies A and B behave most like one another. We implemented a similarity-weighted measure using the pairwise similarity between each pair of animals determined by an independent triadic comparison task (Lee, Abramyan, & Shankle., 2015; Westfall & Lee, 2020). Using this extension of the metric changes the distances between A and B to 30.1, between A and C to 41.7, and between B and C to 26.4, so that B and C become the most similar. Person A breaks the recall of the African zoo animals across extremes of the list, with "elephant" and "giraffe" first and "zebra" and "tiger" last. Persons B and C, in contrast, recall these animals near each other, although not in the same order as one another. The similarity weighting gives less penalty to the transposition of semantically-related animal names, which leads to B and C being measured as having given the most similar responses.

We again use MDS visualizations to explore the overall relationships between people's free recall patterns, based on the Kendall distance measures. The left-hand panels of Figure 3 show the visualizations for the unweighted metric, in the top panel for 15 labeled people, including A–C above, and in the bottom panel for all 200 people. The right-hand panels show the corresponding visualizations for the similarity-weighted metric. It is clear that the inclusion of similarity information leads to more clustering between the recall patterns, suggesting the presence of different recall patterns that can be understood in terms of the semantic relationships between the stimuli being recalled.

Top-k: Expert Sporting Predictions

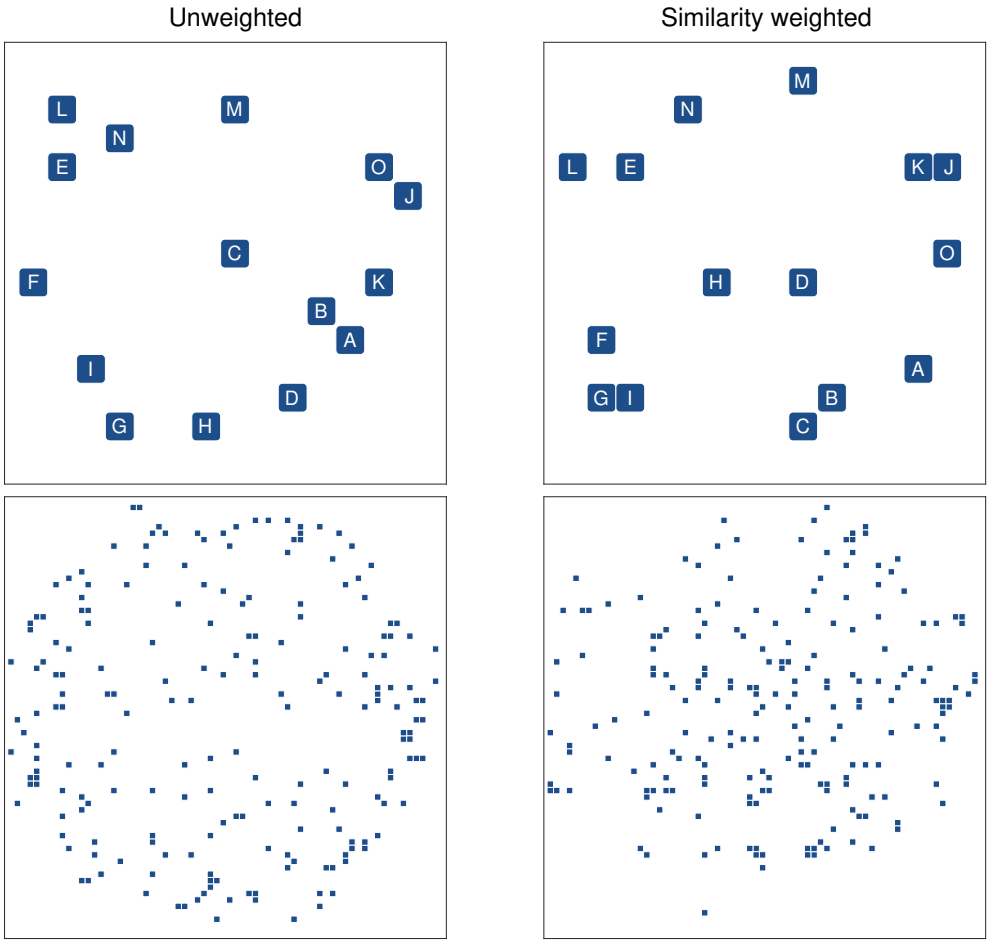
Our last application involves predictions about player performance for the 2017 American Football season by experts from the fantasy football website fantasypros.com.[6] On the website, experts provide rankings each week for each playing position commonly used in fantasy football. These rankings serve as advice for players as to which players they should place in their fantasy teams each week. We focus on the rankings of all 85 experts, but just for week 10 of the season, and just for the "kicker" position. We chose the kicker position because it is the one for which different experts often rank different numbers of players. In week 10, experts ranked between 13 and 20 kickers, with a median of 19. Since a typical fantasy league has around ten players, each of whom own one or two kickers, it is likely that even the ranking of kickers near the end of the list is relevant to some players in the league.

Table 3 shows the actual points earned by each kicker[7], as well as the ranking provided by two of the experts. The Kendall distance provides a natural way of measuring the performance of the experts, by quantifying how close their predictions are to the truth. Some players scored the same number of points, which leads to ties in the true ranking. This can be accommodated using similarity weights, assigning a weight of 0 to any pair of kickers who are tied, and 1 to any pair of players who are not tied.

In addition, because the experts ranked different num-

[6] For those unfamiliar with fantasy football, we suggest the Wikipedia entry at https://en.wikipedia.org/wiki/Fantasy_football_(gridiron)
[7] Players included in expert predictions but not listed did not score any points.

Figure 3 ■ MDS visualization of the similarities between recall patterns of animal names based on the unweighted Kendall distance (left panels) and similarity-weighted Kendall distance (right panels). The top panels show 15 labeled people, while the bottom panels show all 200 people. The distances between the squares/points represent their similarities according to the Kendall distance.



bers of players, their Kendall distance depends on the setting of the missingness parameter. For the optimistic setting ($z = 0$), expert A has accuracy 103 and expert B has accuracy 133, so that expert A is measured as having made better predictions. For the neutral setting ($z = \frac{1}{2}$), expert A has accuracy 153.5 and expert B has accuracy 149.5, so they are very similar. For the pessimistic setting ($z = 1$), expert A has accuracy 204 and expert B has accuracy 167, so now expert B is measured as having made better predictions.

In this application, the optimistic setting seems inappropriate. Expert A included only 14 players in their ranking, whereas Expert B included 20 players. Setting $z = 0$ means that whenever two players (e.g., Mason Crosby and Blair Walsh) are not ranked by an expert, this expert is given the benefit of the doubt and is not penalized. Expert B does include these two players, but predicts their ranking incorrectly, and is penalized for it. This property makes it appealing for an expert to only include the few players that they are very sure about, which is not what is sought from a good prediction. Both the neutral and pessimistic settings seem more appropriate, since they penalize experts who fail to make predictions about players.

Figure 4 shows the change in the Kendall distance for optimistic, neutral, and pessimistic top-k measures for all of the experts. Experts are represented by jittered markers with lines connecting the same expert under each measure. Increasing pessimism leads to the experts who ranked fewer players being penalized more heavily for these missing data. Thus, while the distance measure increases for all of the experts as pessimism increases, it in-

Table 3 ■ The number of fantasy points scored by kickers in week 10 of the 2017 American National Football League season, the true ranking of the players according to these point totals, and the ranked predictions of two experts from fantasypros.com.

Points	True ranking	Expert A	Expert B
15	Greg Zuerlein	Greg Zuerlein	Stephen Gostkowski
12	Nick Rose	Matt Bryant	Greg Zuerlein
11	Mason Crosby	Stephen Gostkowski	Matt Bryant
11	Stephen Gostkowski	Matt Prater	Matt Prater
11	Wil Lutz	Josh Lambo	Mike Nugent
10	Connor Barth	Graham Gano	Ryan Succop
10	Brandon McManus	Chris Boswell	Chris Boswell
9	Matt Bryant	Kai Forbath	Chandler Catanzaro
9	Graham Gano	Blair Walsh	Kai Forbath
9	Patrick Murray	Ryan Succop	Steven Hauschka
8	Kai Forbath	Wil Lutz	Wil Lutz
8	Matt Prater	Steven Hauschka	Brandon McManus
8	Blair Walsh	Mike Nugent	Josh Lambo
7	Robbie Gould	Chandler Catanzaro	Adam Vinatieri
7	Aldrick Rosas		Graham Gano
6	Chris Boswell		Blair Walsh
6	Zane Gonzalez		Robbie Gould
6	Josh Lambo		Mason Crosby
6	Ryan Succop		Connor Barth
5	Nick Novak		Nick Rose

creases more quickly for some experts.

Conclusion

In this article, we aimed to discuss three extensions of the Kendall distance metric that are useful for analyzing ranking data in psychological research, as well as demonstrating the ability of the metric to accommodate top-k lists. Our applications gave worked examples of how the extensions can help improve the measurement of key properties of ranking data in the context of specific research goals. Two of the applications focused on measuring people's accuracy, and two focused on measuring the extent and nature of the individual differences between people. Measuring performance and individual differences are among the most common and basic goals of data analysis in psychology.

While we mostly applied the extensions separately, the second and fourth applications showed that multiple extensions can be used simultaneously. There is nothing preventing Kendall distance measures being designed to be sensitive to items, their positions, and their similarities in top-k lists where different people have different k. This underscores the flexibility and generality of the metric, and its ability to be adapted to answer specific questions in specific research contexts. While this flexibility should help improve data analysis, it may be important to use preregistration to make a clear whether and how the extensions to the metric are used in an exploratory way (Lee et al., 2019). Future directions for the weighted Kendall distance can focus on developing a hypothesis testing framework for the distance, and to apply the weighted metric to Mallow's ϕ model (Mallows, 1957) for finding the modal ranking (e.g., Chierichetti, Dasgupta, Haddadan, Kumar, & Lattanzi, 2018).

Open Practices Statement

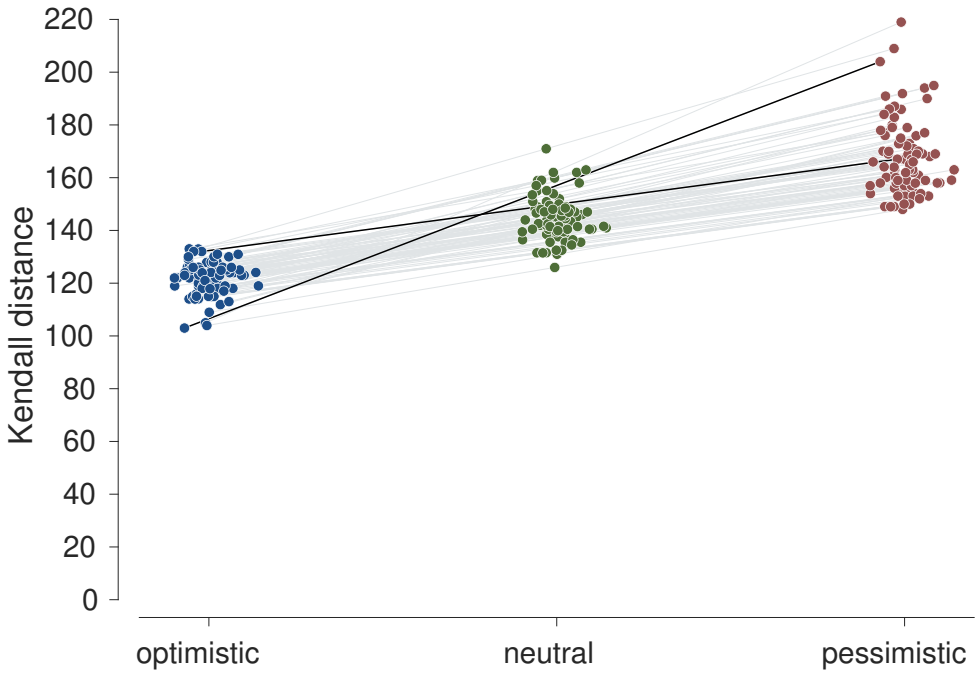
An OSF project page associated with this article is available at https://osf.io/6k9t8/ It includes the R script used for calculating the weighted Kendall distance and example applications of the toy example described in this article. The code is also available on https://github.com/JohnnyDoorn/KendallWeightedDistance.

References

Adomavicius, G., Bockstedt, J., & Curley, S. P. (2015). Bundling effects on variety seeking for digital information goods. *Journal of Management Information Systems*, *31*, 182–212.

Altmann, E. M. (2003). Reconstructing the serial order of events: A case study of September 11, 2001. *Applied*

Figure 4 ■ Optimistic, neutral, and pessimistic Kendall distance measures of accuracy for 85 experts from fantasypros.com. Each expert predicted the fantasy football performance of kickers in week 10 of the American National Football League 2017 season. Each point corresponds to a participant, jittered around the Kendall distance measure. The same participant for each measure is connected by a gray line. Experts A and B from Table 3 are highlighted by black lines.



Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition, *17*, 1067–1080. doi:10.1002/acp.986

Beg, M. S., & Ahmad, N. (2003). Soft computing techniques for rank aggregation on the world wide web. *World Wide Web*, *6*, 5–22. doi:10.1023/A:1022344031752

Borg, I., & Groenen, P. (1997). *Modern multidimensional scaling: Theory and applications*. New York: Springer-Verlag.

Bousfield, W. A. (1953). The occurrence of clustering in the recall of randomly arranged associates. *The Journal of General Psychology*, *49*, 229–240. doi:10.1080/00221309.1953.9710088

Bousfield, W. A., & Sedgewick, C. H. W. (1944). An analysis of sequences of restricted associative responses. *The Journal of General Psychology*, *30*, 149–165. doi:10.1080/00221309.1944.10544467

Brandt, F., Conitzer, V., Endriss, U., Lang, J., & Procaccia, A. D. (2016). *Handbook of computational social choice*. Cambridge University Press.

Chierichetti, F., Dasgupta, A., Haddadan, S., Kumar, R., & Lattanzi, S. (2018). Mallows models for top-k lists. *Advances in Neural Information Processing Systems*, *31*, 4382–4392.

Fagin, R., Kumar, R., & Sivakumar, D. (2003). Comparing top k lists. *SIAM Journal on Discrete Mathematics*, *17*, 134–160. doi:10.1137/S0895480102412856

Gordon, K. H. (1924). Group judgments in the field of lifted weights. *Journal of Experimental Psychology*, *7*, 398–400.

Hamada, D., Nakayama, M., & Saiki, J. (2020). Wisdom of crowds and collective decision-making in a survival situation with complex information integration. *Cognitive Research: Principles and Implications*, *5*, 1–15.

Healey, M. K., & Kahana, M. J. (2014). Is memory search governed by universal principles or idiosyncratic strategies? *Journal of Experimental Psychology: General*, *143*, 575.

Järvelin, K., & Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, *20*, 422–446.

Kendall, M. (1938). A new measure of rank correlation. *Biometrika*, *30*, 81–93. doi:10.2307/2332226

Kumar, R., & Vassilvitskii, S. (2010). Generalized distances between rankings. In *Proceedings of the 19th inter-*

national conference on world wide web (pp. 571–580). ACM.

Lee, M. D., Abramyan, M., & Shankle., W. R. (2015). New methods, measures, and models for analyzing memory impairment using triadic comparisons. *Behavior Research Methods*, *48*, 1492–1507. doi:10.3758/s13428-015-0662-4

Lee, M. D., Steyvers, M., & Miller, B. J. (2014). A cognitive model for aggregating people's rankings. *PLoS ONE*, *9*, 1–9. doi:10.1371/journal.pone.0096431

Lee, M. D., Criss, A. H., Devezer, B., Donkin, C., Etz, A., Leite, F. P., ... White, C. N., et al. (2019). Robust modeling in cognitive science. *Computational Brain & Behavior*, *2*, 141–153. doi:10.1007/s42113-019-00029-y

Mallows, C. L. (1957). Non-null ranking models. *Biometrika*, *44*, 114–130.

Milosavljevic, M., Navalpakkam, V., Koch, C., & Rangel, A. (2012). Relative visual saliency differences induce sizable bias in consumer choice. *Journal of Consumer Psychology*, *22*, 67–74.

Romney, A. K., Brewer, D. D., & Batchelder, W. H. (1993). Predicting clustering from semantic structure. *Psychological Science*, *4*, 28–34. doi:10.1111/j.1467-9280.1993.tb00552.x

Selker, R., Lee, M. D., & Iyer, R. (2017). Thurstonian cognitive models for aggregating top-n lists. *Decision*, *4*, 87–101. doi:10.1037/dec0000056

Shankle, W. R., Mangrola, T., Chan, T., & Hara, J. (2009). Development and validation of the Memory Performance Index: Reducing measurement error in recall tests. *Alzheimer's & Dementia*, *5*, 295–306. doi:10.1016/j.jalz.2008.11.001

Shaw, C. J., & Trimble, T. (1963). Algorithm 175: Shuttle sort. *Communications of the ACM*, *6*, 312–313.

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology*, *15*, 72–101. doi:10.1037/11491-005

Westfall, H. A., & Lee, M. D. (2020). A model-based analysis of the impairment of semantic memory. *Manuscript submitted for publication*.

Open practices

The *Open Material* badge was earned because supplementary material(s) are available on osf.io/6k9t8/.

Citation

van Doorn, J., Westfall, H. A., & Lee, M. D. (2021). Using the weighted Kendall Distance to analyze rank data in psychology. *The Quantitative Methods for Psychology*, *17*(2), 154–165. doi:10.20982/tqmp.17.2.p154

Received: 31/03/2021 ~ Accepted: 14/06/2021