Examination of the similarity between a new Sigmoid function-based consensus ranking method and four commonly-used algorithms

Madjid Tavana*

Professor, Management Information Systems,
La Salle University,
1900 West Olney Avenue,
Philadelphia, PA 19141, USA
E-mail: tavana@lasalle.edu
URL: http://lasalle.edu/~tavana
*Corresponding author

Frank LoPinto

Assistant Professor, Management Information Systems,
La Salle University,
1900 West Olney Avenue,
Philadelphia, PA 19141, USA
E-mail: lopinto@lasalle.edu
URL: http://lasalle.edu/~lopinto

James W. Smither

Professor, Management,
La Salle University,
1900 West Olney Avenue,
Philadelphia, PA 19141, USA
E-mail: smither@lasalle.edu
URL: http://lasalle.edu/~smither

Abstract: The problem of aggregating individual rankings to create an overall consensus ranking representative of the group is of longstanding interest in group decision making. The problem arises in situations where a group of *k* Decision Makers (DMs) are asked to rank order *n* alternatives. The question is how to combine the DMs' rankings into one consensus ranking. Several different approaches have been suggested to aggregate DM responses into a compromise or consensus ranking, however, none is generally recognised as being the best and the similarity of consensus rankings generated by these algorithms is largely unknown. In this paper, we propose a new Weighted-sum ordinal Consensus ranking Method (WCM) with the weights derived from a Sigmoid function. We run Monte Carlo simulation across a range of *k* and *n* to compare the similarity of the consensus rankings generated by our method with the best-known method of Borda–Kendall (BAK; Kendall, M. (1962) *Rank correlation methods*. New York, NY: Hafner) and two other commonly used techniques proposed by Beck, M.P. and Lin, B.W. (1983) 'Some heuristics for the consensus ranking problem', *Computers and Operations Research*, Vol. 10, pp.1–7 and Cook, W.D. and Kress, M. (1985) 'Ordinal rankings with intensity of preference', *Management Science*, Vol. 31, pp.26–32. WCM and BAK yielded the most similar consensus rankings (mean *tau-x* = .91). As the number of alternatives to be ranked increased, the similarity of rankings generated by the four algorithms decreased. Although consensus rankings generated by different algorithms were similar, differences in rankings among the algorithms were of sufficient magnitude that they often cannot be viewed as interchangeable from a practical perspective.

Keywords: consensus; group decisions; Monte Carlo; ordinal; preference; ranking; voting.

Reference to this paper should be made as follows: Tavana, M., LoPinto, F. and Smither, J.W. (2008) 'Examination of the similarity between a new Sigmoid function-based consensus ranking method and four commonly-used algorithms', *Int. J. Operational Research*, Vol. 3, No. 4, pp.384–398.

Biographical notes: Madjid Tavana is a Professor of Management Information Systems and the Lindback Distinguished Chair of Information Systems at La Salle University, where he also served as the Chairman of the Management Department and Director of the Center for Technology and Management. He has been a Faculty Fellow in Aeronautics and Space Research at NASA's Kennedy Space Center, developing Group Decision Support Systems for the Space Shuttle Program and Johnson Space Center, developing Intelligent Decision Support Systems for Human Exploration of Mars at the Mission Control Center. In 2005, he was awarded the prestigious Space Act Award by NASA. He holds an MBA, PMIS, and PhD in Management Information Systems and received his Post-Doctoral Diploma in Strategic Information Systems from the Wharton School of the University of Pennsylvania. He has published in such journals as *Decision Sciences, Interfaces, Information and Management, Computers and Operations Research*, and *Journal of the Operational Research Society*, among others.

Frank LoPinto is an Assistant Professor of Management Information Systems at La Salle University. He received his Bachelor of Civil Engineering from University of Delaware and has an MBA from University of North Carolina at Greensboro, and a PhD in Business Administration (Business Information Technology) from Virginia Polytechnic Institute.

James W. Smither is Lindback Professor of Human Resource Management at La Salle University. He received his BA (in psychology) from La Salle and has an MA from Seton Hall University, an MA from Montclair State University, and a PhD in industrial/organizational psychology from Stevens Institute of Technology. He has published over 45 scholarly articles and chapters; his research has been published in journals such as *Personnel Psychology*, *Journal of Applied Psychology*, and *Organizational Behavior and Human Decision Processes*. He is currently a Consulting Editor at the *Journal of Applied Psychology*. From 1997 to 2003, he served as an Associate Editor of Personnel Psychology. He is a Fellow of the Association for Psychological Science and the Society for Industrial, and Organizational Psychology.

1 Introduction

The problem of aggregating individual rankings to create an overall consensus ranking representative of the group is of longstanding interest in many interdisciplinary fields like organisational sciences, psychology, public policy administration, marketing research, and management science. In particular, the problem of consensus ranking when the individual preferences are represented as ordinal ranking vectors has been investigated extensively by many researchers. Ordinal preferences are commonly used when the information available to the Decision Maker (DM) is of such a nature that only an expression of preference (not degree of preference) can be given. The attractiveness of ordinal representation is due in great part to the minimal amount of information required, i.e. each DM need to express a preference of one alternative over another – not the degree of preference required with cardinal consensus ranking approaches.

1.1 General background for the research

Numerous integer programming, goal programming, and multiple criteria approaches have been proposed to aggregate DM responses into a compromise or consensus ranking, since the early work of Borda (1781), Black (1948), Arrow (1951), Kemeny and Snell (1962), Kendall (1962), Inada (1969), Bogart (1973, 1975), Bowman and Colantoni (1973), and Blin and Whinston (1974). However, none is generally recognised as being the best and the similarity of consensus rankings generated by these algorithms is largely unknown.

Nash (1953), Harsanyi (1955), DeGroot (1974), Keeney and Kirkwood (1975), Bodily (1979) and Brock (1980) have suggested various classical consensus ranking methods requiring explicit knowledge of the DMs' utility functions and explicit interpersonal comparisons of utility. The difficulty of estimating an individual's utility function and explicit interpersonal utility comparisons has severely limited practical applications of such methods.

The simplest form of group consensus, the majority rule, is used when individual preferences are presented as ordinal ranking vectors. Borda (1781) proposed the 'Method of Marks' to derive a consensus of opinions by determining the average of the ranks assigned by DMs to each alternative, with the winning alternative being the one with the lowest average. A similar version of this model was later presented by Kendal (1962). Kendall (1962) was the first to study the ordinal ranking problem in a statistical framework by approaching the problem as an estimation problem. Kendall's solution – to rank alternatives according to the sums of the ranks – is precisely Borda's method of marks. Borda–Kendall (BAK) technique is the most widely used consensus ranking method in practice because of its computational simplicity (Cook et al., 1997; Jensen, 1986). However, many authors regard the BK technique as *ad hoc* because this method does not satisfy the social welfare axioms of universal domain, transitivity, unanimity, rank reversal, and non-dictatorship proposed by Arrow (1951). Cook and Seiford (1982) further studied the BAK technique and proposed a 'minimum variance' method for determining the consensus ranking. Inada (1969) also studied the majority rule and showed the conditions under which the majority rule satisfies Arrow's axioms.

Another popular method for deriving a group consensus is to define a distance function for all rankings and then finding the ordinal ranking with the minimum distance. Kemeny and Snell (1962) proposed a distance measure that represented the degree of correlation between a pair of rankings along with a set of axioms similar to those given by Arrow (1951) and showed that the median and mean rankings are acceptable forms of consensus. Bogart (1973, 1975) considered both transitive and intransitive orderings and generalised the Kemeny and Snell (1962) theory. Cook and Saipe (1976) developed a branch and bound algorithm to determine the median of a set of ordinal rankings. Cook and Seiford (1978) examined the problem of deriving a consensus among a set of ordinal rankings and developed a set of axioms for distance measures. Cook and Kress (1985) extended Kemeny and Snell (1962) theory by suggesting a method for representing strength of preference within an ordinal scale. Cook and Kress (1992), Cook et al. (1996, 2007) and Cook (2006) further studied the complete preference case using distance functions.

Bowman and Colantoni (1973) and Blin and Whinston (1974) presented integer programming models to solve the majority rule problems. Cook and Kress (1985) developed a network model for deriving the optimal consensus ranking that minimises disagreement among a group of DMs. Ali, Cook and Kress (1986) presented an integer programming approach for consensus ranking. Cook et al. (2005) compared integer programming with heuristics models. Fan, Gordon and Pathak (2004) proposed a generic ranking function and genetic programming for group consensus ranking. Iz and Jelassi (1990) proposed goal programming for formulating and solving consensus ranking problems. Gonzalez-Pachon and Romero (2001) followed up by developing an interval goal programming model for aggregating incomplete individual preferences in group consensus problems. Cook and Kress (1991) proposed an ordinal ranking model for on the basis of multiple criteria. Muralidhara, Anatharaman and Deshmukh (2002) proposed a Consensus Ranking Model (CRM) combining ordinal rankings with multiple criteria and importance weights. Tavares (2004) proposed CRM in terms of the weights space by avoiding any assumptions about the distance between ranked alternatives. Leyva-Lopez and Fernandez-Gonzalez (2003) and Kengpol and Tuominen (2006) developed decision support frameworks for group consensus ranking. Jensen (1986) compared a few of the most widely used CRM and suggested that various methods proposed in the literature yield controversial rankings and/or rankings which are vulnerable to considerable dispute.

1.2 Specific background for the research

Beck and Lin (1983) have developed a procedure, known as Maximise Agreement Heuristic (MAH), for approximating the optimal consensus rankings. MAH is commonly used in practice because of its simplicity, flexibility, and general performance (Lewis and Butler, 1993; Tavana, Kennedy and Joglekar, 1996; Tavana, 2002, 2003; Kengpol and Tuominen, 2006). MAH is intended to derive consensus orderings that reflect collective DM agreement. An agreement is defined as a case where alternative i is preferred to alternative j by a given DM and alternative i is ranked above alternative j in the final consensus ranking. A disagreement is a case where alternative i is preferred to alternative j by a DM and alternative j is ranked above alternative i in the final consensus ranking. Under the MAH, alternative i is positioned above alternative j in the final ordering if the difference in total DM agreement and disagreement about the relative orderings of alternatives i and j is positive, and alternative j is positioned above alternative i if this difference is negative. These positioning assignments are iteratively made based on the maximum absolute agreement/disagreement difference of all unassigned alternatives. A detailed explanation of MAH is presented in Appendix A.

Cook and Kress (1985) have suggested a more complicated method referred to as CRM for representing strength of preference within an ordinal scale. In CRM, DM orders n alternatives in q positions where $n \leq q$. The resulting ranking shows the DM's order of preference and the relative positioning of the alternatives represents his or her intensity of preference. CRM guarantees common units across DMs if each DM orders the same number of alternatives into the same number of positions. In order to minimise the number of iterations in our Monte Carlo simulation, ties are not considered in CRM ($n = q$). A detailed explanation of CRM is presented in Appendix B.

The remainder of the paper is organised as follows. Section 2 describes the mathematical details of our model followed by a description of our study in Section 3. In Section 4, we discuss our results and in Section 5, we present our conclusions.

2 The weighted-sum model

In this paper, we propose a new weighted-sum ordinal consensus ranking method referred as Weighted-sum Consensus Model (WCM) with the weights derived from a Sigmoid function. First, we identify an initial preference matrix, R, where the ith row corresponds to the ith DM and the jth column corresponds to the jth ranking. The element in the ith row and the jth column, r_{ij}, shows the number of times alternative i is ranked 1st, 2nd, …, and nth. Next, ranking of r's are weighted according to the Sigmoid function. The Sigmoid function produces an 'S-shaped' curve with larger weights for extreme cases (top and bottom alternatives) and smaller weights for the centre (middle alternatives). A detailed discussion of Sigmoid function properties can be found in Grossberg (1973) and Menon et al. (1996). When discussing the process of rank ordering candidates (employees or applicants), Guion (1998) notes that extreme judgments (e.g. distinguishing the best from the worst employees) are easy, but differences near the centre of the distribution are harder to identify (p.545). This observation is consistent with an analysis of rankings from Olympic figure skating competitions where Truchon (2004) found that most cycles (i.e. intransitivities) involved middle ranked skaters. Both of these observations are consistent with the longstanding finding from perception and psychophysics research that judgment time is inversely related to the distance between the stimuli being judged (e.g. Dashiell, 1937). The Sigmoid function used in our weighted-sum model ensures exponentially differentiable weights among the top, bottom, and middle rankings by providing heavier weights to the extreme judgments where the best and the worst candidates are more distinguishable than the middle candidates. We also examine the similarity in rankings generated by our method with the best-known method of BAK and two other commonly used techniques proposed by Beck and Lin (1983) and Cook and Kress (1985). We use Monte Carlo simulation to examine the extent to which these algorithms yield similar rank ordering across a range of k and n.

To formulate an algebraic model of WCM, let us consider a general consensus ranking problem with k DMs and n alternatives. Let us assume r_{ij} is the alternative number of the ith DM on jth ranking ($i = 1, ..., k$ and $j = 1, ..., n$) and $R=(r_{ij})$ represents the matrix where the ith row corresponds to the ith DM, the jth column corresponds to the jth ranking, and the element in the ith row and the jth column, r_{ij}, is the alternative that the ith DM put into the jth ranking.

Let us further assume f_{mj} is the number of times alternative *m* appears in column *j* of matrix *R* ($m = 1, ..., n$ and $j = 1, ..., n$) and $F=(f_{mj})$ represent the matrix where the *m*th row corresponds to the *m*th alternative, the *j*th column corresponds to the *j*th ranking, and the element in the *m*th row and the *j*th column, f_{mj}, is the number of times that alternative *m* is placed in the *j*th ranking. We maximise w_i, the total weighted-sum score for alternative $i\left(\underset{i=1,\ldots,k}{Max.w_i} = \sum_{\substack{j=1 \\ i=1,\ldots,k}}^{n} f_{mj} \,.\, S_j\right)$; where S_j is the weight of the *j*th ranking using the Sigmoid function $\left(S_j = \frac{1}{1+e^{-x}}\right)$; *x* is the linear interpolation bounded by $\pm\alpha$; and α is the linear limiting control used to ensure exponentially differentiable weights among the top, bottom, and middle rankings.

3 The study and results

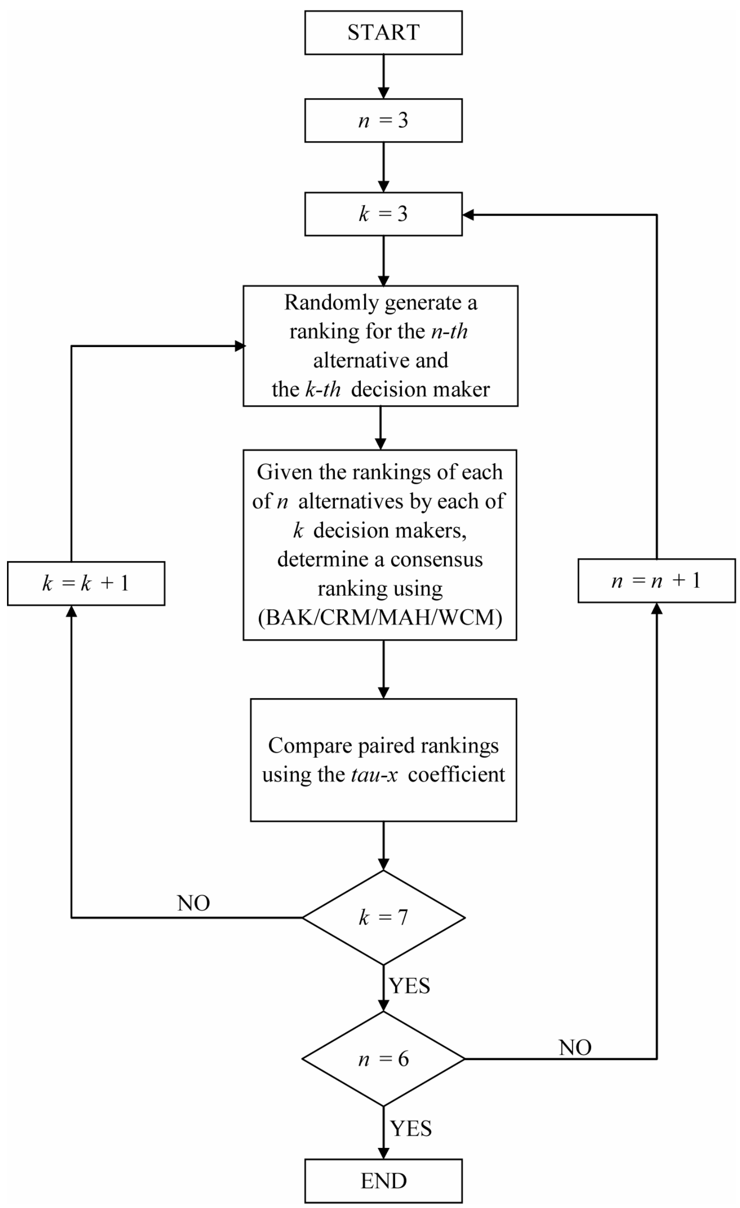
In this study, we used Monte Carlo simulation (See Figure 1) to compare the performance of our method WCM with BAK (Kendall, 1962), MAH (Beck and Lin, 1983), and CRM (Cook and Kress, 1985). Our testing platform was a Pentium 4 CPU, 3.33 GHz, with 1.00 GB RAM, running under Microsoft Windows XP.

We randomly generated rankings of *n* alternatives for each of the *k* DMs. Initial individual rankings for each trial were generated using uniformly distributed random numbers from the Mersenne Twister (Matsumoto and Nishimura, 1998) random number generator. Each of the four consensus algorithms was then used to aggregate the individual rankings into a single consensus ranking of all *n* items for all *k* DMs.

For each scenario, the number of alternatives to be ranked, *n*, varied from 3 to 6. The number of DMs, *k*, varied from 3 to 7. One thousand repetitions for each *n*,*k* combination were conducted. The result of the experiment was 20 separate unique *n,k* combinations. One thousand repetitions of each unique *n,k* combination resulted in 20,000 total trials. For each trial, each of the four techniques was used to generate group consensus rankings. Therefore, the analysis generated a total of 80,000 data vectors in the form of group consensus rankings.

For each trial, where each trial begins with the same initial individual DM rankings of the *n* alternatives, there were four consensus rankings of the alternatives (WCM, BAK, MAH, and CRM). One goal of our Monte Carlo simulation was to examine the extent to which the four algorithms yield similar rank ordering across a range of problems with *k* DMs and *n* alternatives. The most commonly used measures of association (similarity) with ordinal rankings are Spearman's *rho* and Kendall's *tau*. *Rho* and *tau* are not identical in magnitude because their underlying logic and computational formulae are quite different. The choice of measures is not a trivial one because Kendall (1938) has noted that values of *tau* and *rho* are similar at some magnitudes, but differ appreciably at others. Indeed, Strahan (1982) and Gibbons (1976) have noted that, for most degrees of association that occur in practice (i.e. when absolute values are not too close to 1), *tau* is typically smaller in absolute value than Spearman's *rho*, with *tau* often no more than two-thirds the size of *rho*. Some advantages of *tau* over *rho* have long been discussed (Schaeffer and Levitt, 1956). For example, Kendall initially noted that the distribution of *tau* is normal not only for large values of *N* (as is *rho*) but also for very small values. Hays (1973) noted that *rho* is, in most instances, a biased estimator, whereas *tau* provides an unbiased estimate of the true population correlation. Noether (1986) notes that *tau* has an intuitively simple interpretation (i.e. the difference between the probability that two alternatives are in the same order vs. the probability that the two alternatives are in a different order) whereas *rho* does not. One study found that, relative to *rho*, *tau* provided adequate control of type I errors and tighter confidence intervals (Arndt and Turvey, 1999).

Figure 1 Monte Carlo simulation flowchart



Despite its advantages relative to *rho*, Emond and Mason (2002) have recently shown that Kendall's *tau* is a flawed measure of agreement between weak orderings (i.e. when tied rankings are allowed). In response, Emond and Mason (2002) presented a new rank correlation coefficient (*tau-x*) and described its application to consensus ranking problems. *Tau-x* is an extension of Kendall's *tau* that handles tied rankings in a different way. It is the unique rank correlation coefficient that is equivalent to the Kemeny–Snell distance metric on the set of all weak orderings of *n* alternatives. Emond and Mason (2002) describe the advantages of *tau-x* (relative to Kendall's *tau*) and show that *tau-x* provides a more mathematically tractable solution because all the ranking information can be summarised in a single combined input matrix. Moreover, *tau-x* allows researchers to handle consensus ranking problems with weights, ties, and partial inputs. In view of the above, we used *tau-x* to assess the similarity of consensus rankings generated by the four different approaches.

For each combination of *k* and *n*, Table 1 reports the mean and standard deviation of the *tau-x* coefficients comparing each of the four consensus algorithms to the other three algorithms. We conducted several analyses help to clarify the results presented in Table 1. First, it is noteworthy that across all 120 cells in Table 1, the mean similarity (*tau-x*) of rankings was 0.82 (SD = 0.18). This indicates that the consensus rankings generated by the four algorithms (WCM, CRM, MAH, and BAK) were somewhat similar.

Table 1 Mean similarity (*tau-x*) of consensus rankings generated by four algorithms for combinations of *n* alternatives ranked by *k* decision makers

		CRM				*MAH*				*WCM*			
		3n	*4n*	*5n*	*6n*	*3n*	*4n*	*5n*	*6n*	*3n*	*4n*	*5n*	*6n*
BAK	3*k*	.94	.82	.76	.73	.94	.90	.88	.87	1.00	.92	.93	.89
	4*k*	.75	.73	.69	.70	.90	.86	.85	.83	1.00	.90	.91	.86
	5*k*	.84	.77	.73	.73	.92	.87	.84	.84	1.00	.88	.90	.85
	6*k*	.76	.73	.73	.72	.89	.86	.84	.83	1.00	.86	.88	.83
	7*k*	.81	.75	.73	.72	.88	.86	.83	.82	1.00	.85	.89	.84
CRM	3*k*					1.00	.90	.84	.80	.94	.81	.76	.71
	4*k*					.81	.76	.71	.71	.75	.71	.68	.66
	5*k*					.92	.85	.79	.77	.84	.75	.72	.69
	6*k*					.81	.77	.75	.74	.76	.71	.71	.68
	7*k*					.92	.82	.78	.75	.81	.70	.71	.68
MAH	3*k*									.94	.87	.85	.82
	4*k*									.90	.79	.80	.75
	5*k*									.92	.80	.80	.77
	6*k*									.89	.77	.78	.74
	7*k*									.88	.76	.79	.75

Using the data from all 120 cells in Table 1, we examined whether the number of alternatives to be ranked (*n*) or the number of DMs (*k*) was related to the similarity of rankings. Across all 120 cells in Table 1, the correlation of *n* with mean similarity (*tau-x*) of rankings was –0.51 ($p < 0.01$). As the number of alternatives to be ranked increased, the similarity of rankings generated by the four algorithms decreased. When there were three alternatives to be ranked, the mean similarity among the four algorithms was 0.89.

When there were four alternatives to be ranked, the mean similarity was 0.81. When there were five alternatives to be ranked, the mean similarity was 0.80. When there were six alternatives to be ranked, the mean similarity was 0.77. The correlation of k with mean similarity (*tau-x*) of rankings was –0.21 ($p < 0.05$). That is, as the number of DMs increased, there was a slight tendency for the mean similarity of rankings to decrease.

Next, to determine whether some pairs of algorithms generated more similar consensus rankings than other pairs, we conducted a one-way analysis of variance (ANOVA) where the dependent variable was the mean similarity (*tau-x*) of rankings and the independent variable was the six pairs of algorithms (i.e. WCM and MAH, WCM and CRM, WCM and BAK, MAH and CRM, MAH and BAK, and CRM and BAK). Results indicated that there were significant differences in the similarity of rankings generated by different pairs of algorithms ($F = 22.85$, df = 5, 114, $p < 0.01$). The mean similarity (*tau-x*) for each pair of algorithms is listed in Table 2.

Table 2 Mean similarity (*tau-x*) for each pair of consensus algorithms

	Mean similarity	*SD*
WCM and MAH [3, 4]	0.82	0.06
WCM and CRM [1]	0.74	0.07
WCM and BAK [5]	0.91	0.06
MAH and CRM [2, 3]	0.81	0.08
MAH and BAK [4, 5]	0.87	0.03
CRM and BAK [1, 2]	0.76	0.06

Note: Homogeneous subsets (based on *post hoc* Tukey HSD tests) share a common superscript.

Post hoc Tukey Honestly Significant Difference (HSD) tests identified five homogeneous subsets of algorithm pairs. WCM and BAK generated the most similar rankings (mean *tau-x* across all combinations of n and k was 0.91). WCM and CRM generated the least similar rankings (mean *tau-x* across all combinations of n and k was 0.74).

It might be helpful to view the similarity in rankings generated by different consensus algorithms from the perspective of psychometric theory. In psychometrics, reliability is often estimated by examining the correlation between two methods of rank ordering individuals (or objects). If two alternative approaches (e.g. two sets of scorers, two ‘parallel’ forms of a test) yield highly similar (correlated) results, this is viewed as evidence of their reliability. That is, when two scorers generate highly similar ratings (or rankings) of a group of individuals or objects, their ratings (or rankings) are said to be reliable. If two ‘parallel’ forms of a test generate highly similar rank orderings of individuals, they are said to be reliable (i.e. an individual can be expected to obtain nearly the same score regardless of which form of the test the individual is administered). In his classic text, Nunnally (1978) notes that the level of ‘satisfactory’ reliability depends upon how a measure is used. For basic research, Nunnally states that measures with reliability of 0.80 (e.g. the correlation between two ‘parallel’ forms of a test) are useful. In applied settings where an exact score on a measure determines whether an applicant will be accepted into a school, programme, or organisation, higher levels of reliability (above 0.90) are desirable. In this context, it is noteworthy that the similarity of rankings (*tau-x*) generated by WCM and BAK (0.91) exceeds the level associated with satisfactory reliability in applied settings. The similarity of rankings generated by MAH and each of the other three algorithms (BAK, CRM, and WCM) exceeds 0.80, thereby indicating that rankings generated by MAH are quite similar to (but certainly not identical with) rankings generated by the other algorithms. In contrast, the rankings generated by CRM and WCM and by CRM and BAK (0.74 and 0.76, respectively) are similar but cannot be viewed as interchangeable from a practical perspective.

The similarity (or lack of similarity) of consensus rankings generated by different algorithms (or rules) has consequences in applied settings. For example, Truchon (2004) has illustrated (with 30 Olympic figure skating competitions between 1976 and 2002) that the choice of a consensus ranking rule or algorithm can have a real impact on the results.

The finding that consensus rankings generated by different algorithms cannot always be treated as interchangeable begs the question, "Which algorithm generates the 'best' or most useful consensus rankings?" A variety of criteria could be used to evaluate consensus algorithms. Some criteria might focus on the technical properties of the algorithms. For example, algorithms with relatively less computational complexity (e.g. especially when given a large number of DMs or alternatives to be ranked) might be viewed as preferable. Algorithms that minimise the effect of an outlier DM on the final consensus rankings might also be preferred. This perspective might also consider how effectively an algorithm deals with the intransitivities that often emerge when DMs' rankings are generated from paired comparisons of alternatives. Algorithms might also be judged by the value of the information they provide beyond the final consensus ranking of alternatives (e.g. whether the algorithm yields easily interpretable information about how far apart alternatives are in the eyes of DMs). Another criterion could be whether the underlying assumptions of an algorithm appear tenable. For example, BAK implicitly assumes that DMs are able to rank all alternatives from first to last despite findings that such all-at-once comparisons can be quite difficult for many people (Jensen, 1986). Perez and Barba-Romero (1995) proposed three criteria: the frequency/intensity of internal consistencies, the ability to discriminate the best alternatives, and the complexity of the algorithms.

Another criterion could be the ease with which individual DMs (or judges) can manipulate the final consensus ranking. Truchon (2004) provides examples where individual judges sought to manipulate (via insincere voting rather than merely biased voting) the final consensus ranking. Moreover, a judge could be tempted to give the lowest rank to the objectively best alternative in order to enhance the relative standing of his or her preferred alternative in the final consensus ranking. Truchon notes that some ranking rules might be more prone to manipulation than others and that some ranking rules might require more sophistication than others in order to be manipulated successfully (e.g. in some cases manipulation by individuals or small groups might be possible, whereas in other cases manipulation might require large coalitions).

4 Conclusions and future research directions

Research is needed to better understand, how different algorithms perform under different conditions. For example, when there is relatively high agreement among the rankings initially provided by different DMs, consensus controversies are likely to be minimal and a variety of algorithms will lead to similar consensus rankings (Truchon, 2004). Consensus controversies are more likely to arise when there is considerable disagreement among the rankings initially provided by different DMs (especially when these disagreements occur about the top or bottom alternatives). Similarly, consensus controversies are especially likely to arise when there are two or more subsets of DMs, each of whom displays high agreement within the subset but high disagreement with the other subset(s) of DMs. It would therefore be helpful for future research to examine the similarity of consensus rankings generated by different algorithms under such conditions.

Little is knew about how DMs react to different consensus algorithms. For an algorithm to be useful, DMs must ultimately be willing to accept the rankings it generates from their individual rankings. But some algorithms are likely to be especially difficult to explain to DMs (who might have little interest in or understanding of the underlying mathematics). When this occurs, how can DMs, as Jensen (1986) suggested, agree on the consensus scoring rule prior to completing their initial individual rankings? This suggests that it would be helpful to understand DMs' reactions to different algorithms.

It would be especially helpful to understand whether any consensus algorithms closely mirror the consensus decisions reached by interacting groups whose members have first completed individual rankings. That is, individuals sometimes rank alternatives (e.g. job applicants, requests for capital funding) before meeting as a group to discuss their rankings and arrive at a group consensus. Little is understood about how groups ultimately arrive at consensus or the factors that groups consider when determining whether a proposed consensus ranking is acceptable to the group. Group members might begin by debating the merits of their individual rankings. One or more group members might shift their views as new opinions and information are shared among group members. Of course, group members with some characteristics (e.g. high status and assertive) might be more influential than others (e.g. with less status and deferential), and thereby ultimately persuade other group members to accept their rankings. In such cases, the group's consensus ranking will closely correspond to the individual rankings initially generated by some (influential) members but will differ from the individual rankings initially generated by other members. It would be of practical and theoretical value to understand the similarity between consensus rankings arrived at by group discussion vs. consensus rankings arrived at solely by the application of algorithms such as those examined in this paper. If the consensus rankings reached by interacting groups closely resemble those generated by some algorithms (e.g. those that minimise disagreement or distance among a set of DMs) but not other algorithms (those that maximise agreement among a set of DMs or those that consider multiple criteria and importance weights), this would offer insight about how interacting groups ultimately reach consensus. From a practical perspective, if one or more algorithms closely parallel the consensus decisions reached by interacting groups, then groups might be more comfortable using those algorithms to combine their individual rankings without investing the sometimes substantial time and energy required to discuss individual rankings and reach consensus.

As Jensen (1986) noted, no consensus algorithm is likely to be free from criticism in all circumstances. Research that

a examines the performance and properties of algorithms under real-world conditions where consensus controversies are likely to arise

b explores user reactions to and acceptance of different algorithms will enhance the likelihood that at least some consensus algorithms will be adopted by DMs in applied settings.

Acknowledgements

The authors would like to thank the anonymous reviewers and the editor for their constructive suggestions and clear directions.

References

Ali, L., Cook, W.D. and Kress, M. (1986) 'Ordinal ranking and intensity of preference: a linear programming approach', *Management Science*, Vol. 32, pp.1642–1647.

Arndt, S. and Turvey, C. (1999) 'Correlating and predicting psychiatric symptom ratings: Spearman's r versus Kendall's tau correlation', *Journal of Psychiatric Research,* Vol. 33, pp.97–104.

Arrow, K.J. (1951) *Social Choice and Individual Values*. New York, NY: Wiley.

Beck, M.P. and Lin, B.W. (1983) 'Some heuristics for the consensus ranking problem', *Computers and Operations Research*, Vol. 10, pp.1–7.

Black, D. (1948) 'The decisions of a committee using a special majority rule', *Econometrica*, Vol. 16, pp.245–261.

Blin, J.M. and Whinston, A.B. (1974) 'A note on majority rule under transitivity constraints', *Management Science*, Vol. 20, pp.1439–1440.

Bodily, S.E. (1979) 'A delegation process for combining individual utility functions', *Management Science*, Vol. 25, pp.1035–1041.

Bogart, K.P. (1973) 'Preference structures I: distances between transitive preference relations', *Journal of Mathematical Sociology*, Vol. 3, pp.49–67.

Bogart, K.P. (1975) 'Preference structures II: distances between asymmetric relations', *SIAM Journal of Applied Mathematics*, Vol. 29, pp.254–262.

Borda, J.C. (1781) 'Mémoire sur les élections au scrutin', Histoire de l'Académie Royale des Sciences, Paris.

Bowman, V.J. and Colantoni, C.S. (1973) 'Majority rule under transitivity constraints', *Management Science*, Vol. 19, pp.1029–1041.

Brock, H.W. (1980) 'The problem of utility weights in group preference aggregation', *Operations Research*, Vol. 28, pp.176–187.

Cook, W.D. (2006) 'Distance-based and *ad hoc* consensus models in ordinal preference ranking', *European Journal of Operational Research*, Vol. 172, pp.369–385.

Cook, W.D. and Kress, M. (1985) 'Ordinal rankings with intensity of preference', *Management Science*, Vol. 31, pp.26–32.

Cook, W.D. and Kress, M. (1991) 'A multiple criteria decision model with ordinal preference data', *European Journal of Operational Research*, Vol. 54, pp.191–198.

Cook, W.D. and Kress, M. (1992) *Ordinal Information and Preference Structures: Decision Models and Applications*. Englewood Cliffs, NJ: Prentice-Hall.

Cook, W.D. and Saipe, A.L. (1976) 'Committee approach to priority planning: the median ranking method', *Cahiers du Centre d'Études de Recherche Opérationnelle*, Vol. 18, pp.337–351.

Cook, W.D. and Seiford, L.M. (1978) 'Priority ranking and consensus formation', *Management Science*, Vol. 24, pp.1721–1732.

Cook, W.D. and Seiford, L.M. (1982) 'On the Borda–Kendall consensus method for priority ranking', *Management Science*, Vol. 28, pp.621–637.

Cook, W.D., Kress, M. and Seiford, L.M. (1996) 'A general framework for distance-based consensus in ordinal ranking model', *European Journal of Operational Research*, Vol. 92, pp.392–397.

Cook, W.D., Doyle, J., Green, R. and Kress, M. (1997) 'Multiple criteria modeling and ordinal data: evaluation in terms of subset of criteria', *European Journal of Operational Research*, Vol. 98, pp.602–609.

Cook, W.D., Golany, B., Kress, M., Penn, M. and Raviv, T. (2005) 'Optimal allocation of proposals to reviewers to facilitate effective ranking', *Management Science*, Vol. 51, pp.655–661.

Cook, W.D., Golany, B., Penn, M. and Raviv, T. (2007) 'Creating a consensus ranking of proposals from reviewers' partial ordinal rankings', *Computers and Operations Research*, Vol. 34, pp.954–965.

Dashiell, J.F. (1937) 'Affective value-distances as a determinant of esthetic judgment-times', *Am. J. Psychology,* Vol. 50, pp.57–67.

De Groot, M.H. (1974) 'Reaching a consensus', *Journal of American Statistical Association*, Vol. 69, pp.118–121.

Emond, E.J. and Mason, D.W. (2002) 'A new rank correlation coefficient with application to the consensus ranking problem', *Journal of Multicriteria Decision Analysis*, Vol. 11, pp.17–28.

Fan, W., Gordon, M.D. and Pathak, P. (2004) 'A generic ranking function discovery framework by genetic programming for information retrieval', *Information Processing and Management*, Vol. 40, pp.587–602.

Gibbons, J.D. (1976) *Nonparametric methods for quantitative analysis*. New York, NY: Holt, Rinehart and Winston.

Gonzalez-Pachon, J. and Romero, C. (2001) 'Aggregation of partial ordinal rankings: an interval goal programming approach', *Computers and Operations Research*, Vol. 28, pp.827–834.

Grossberg, S. (1973) 'Contour enhancement, short term memory, and constancies in reverberating neural networks', *Studies in Applied Mathematics*, Vol. 52, pp.217–257.

Guion, R.M. (1998) *Assessment, measurement, and prediction for personnel decisions*. Mahwah, NJ: Lawrence Erlbaum.

Harsanyi, J.C. (1955) 'Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility', *Journal of Political Economy*, Vol. 63, pp.309–321.

Hays, W.L. (1973) *Statistics for the social sciences*. New York, NY: Holt, Rinehart, and Winston.

Inada, K. (1969) 'The simple majority rule', *Econometrica*, Vol. 37, pp.490–506.

Iz, P. and Jelassi, M.T. (1990) 'An interactive group decision aid for multiobjective problems: an empirical assessment', *Omega*, Vol. 18, pp.595–604.

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

Keeney, R.L. and Kirkwood, C.W. (1975) 'Group decision making using cardinal social welfare functions', *Management Science*, Vol. 22, pp.430–437.

Kengpol, A. and Tuominen, M. (2006) 'A framework for group decision support systems: an application in the evaluation of information technology for logistics firms', *Int. J. Production Economics*, Vol. 101, pp.159–171.

Kemeny, J.G. and Snell, L.J. (1962) 'Preference ranking: an axiomatic approach', *Mathematical Models in the Social Science*, Chapter 2 (pp.9–23). Boston, MA: Ginn and Co.

Kendall, M.G. (1938) 'A new measure of rank correlation', *Biometrika,* Vol. 30, pp.81–93.

Kendall, M. (1962) *Rank correlation methods*. New York, NY: Hafner.

Lewis, H.S. and Butler, T.W. (1993) 'An interactive framework for multi-person, multiobjective decisions', *Decision Sciences*, Vol. 24, pp.1–22.

Leyva-Lopez , J.C. and Fernandez-Gonzalez, E. (2003) 'A new method for group decision support based on ELECTRE III methodology', *European Journal of Operational Research*, Vol. 148, pp.14–27.

Matsumoto, M. and Nishimura, T. (1998) 'Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator', *ACM Transactions on Modeling and Computer Simulation*, Vol. 8, pp.3–30.

Menon, A., Mehrotra, K., Mohan, C.K. and Ranka, S. (1996) 'Characterization of a class of Sigmoid functions with applications to neural networks', *Neural Networks*, Vol. 9, pp.819–835.

Muralidhara, C., Anatharaman, N. and Deshmukh, S.G. (2002) 'A multi-criteria group decisionmaking model for supplier rating', *The Journal of Supply Chain Management*, Vol. 38, pp.22–33.

Nash, J.F. (1953) 'Two Person Cooperative Games', *Econometrica*, Vol. 21, pp.128–140.

Noether, G.E. (1986.) *Why Kendall tau?* Retrieved June 8, 2006, Available at: http://rsscse.org.uk/ts/bts/noether/text.html

Nunnally, J.C. (1978) *Psychometric theory*. New York, NY: McGraw-Hill.

Perez, J. and Barba-Romero, S. (1995) 'Three practical criteria of comparison among ordinal preference aggregation rules', *European Journal of Operational Research*, Vol. 85, pp.473–487.

Schaeffer, M.S. and Levitt, E.E. (1956) 'Concerning Kendall's tau, a nonparametric correlation coefficient', *Psychological Bulletin*, Vol. 53, pp.338–346.

Strahan, R.F. (1982) 'Assessing magnitude of effect from rank-order correlation coefficients', *Educational and Psychological Measurement,* Vol. 42, pp.763–765.

Tavana, M. (2002) 'Euclid: strategic alternative assessment matrix', *Journal of Multi-Criteria Decision Analysis*, Vol. 11, pp.75–96.

Tavana, M. (2003) 'CROSS: a multicriteria group decision making model for evaluating and prioritizing advanced-technology projects at NASA', *Interfaces*, Vol. 33, pp.40–56.

Tavana, M., Kennedy, D. and Joglekar, P. (1996) 'A group decision support framework for consensus ranking of technical manager candidates', *Omega*, Vol. 24, pp.523–538.

Tavares, L.V. (2004) 'A model to support the search for consensus with conflicting rankings: multitrident', *International Transactions in Operations Research*, Vol. 11, pp.107–115.

Truchon, M. (2004) Aggregation of rankings in figure skating. CIRPEE Working Paper No. 04-14. Available at: http://ssrn.com/abstract=559981.

Appendix A

Maximise Agreement Heuristic Model (MAH) of Beck and Lin (1983)

Assume that each one of a group of k DMs has ranked n alternatives. Assuming further that the opinions of the k DMs are to be valued equally, the MAHR seeks to arrive at the consensus ranking of the alternatives for the group as a whole. According to Beck and Lin, MAHR defines an agreement matrix, A, where each element a_{ij} represents the number of DMs who have preferred alternative i to alternative j. Strict preference is important. If DM is indifferent between i and j, he or she is not counted in $a_{ij.}$ The sum of a_{ij} for each alternative i across all columns represents the positive preference vector, P, where $P_i = \sum_{j=1}^{n} a_{ij}; i = 1, 2, 3, \ldots, n$. Similarly, the sum of a_{ij} for each alternative across all rows represents the negative preference vector, N, where $N_i = \sum_{j=1}^{n} a_{ij}; i = 1, 2, 3, \ldots, n$.

If for alternative i, P_i=0, implying that no DM prefers alternative i to any other alternative, alternative i is placed at the bottom (in subsequent iterations, at the next available position at the bottom) of the final consensus ranking. However, if for alternative i, N_i=0, implying that no DM prefers any other alternative over alternative i, alternative i is placed at the top (in subsequent iterations, at the next available position at the top) of the ranking.

When there are no zero values in either P or N, the difference in total DM agreement and disagreement (P_i–N_i) is calculated for each alternative, and alternative i with the largest absolute difference $|P_i - N_i|$ is considered. If (P_i–N_i) is positive, alternative i is placed in the next available position at top of the final consensus ranking, and if the difference is negative, alternative i is placed in the next available position at the bottom of the consensus ranking. Any ties are broken arbitrarily. Once an alternative is assigned a position in the final consensus ranking, that alternative is eliminated from further consideration. The remaining alternatives form a new matrix and the process is repeated until all alternatives are ranked.

Appendix B

Consensus Ranking Model (CRM) of Cook and Kress (1985)

Assume that a group of k DMs has ranked n alternatives ($v_1,\ldots, v_n$). Each DM ranks the alternatives. A preference intensity matrix, $M = \{m_{ij}\}$, is given for each DM where m_{ij} is the number of times v_i is preferred over v_j ($m_{ij} = 0$ if v_i is tied with v_j). A consensus ranking matrix, $B = \{b_{ij}\}$, is the matrix that minimises the distance between B and the matrices M_i, $M(B) = \min_{B \in D} \sum_{i=1}^{k} d(M_i, B)$ and D is the set of all $n \times n$ preference intensity matrices, $d(M,B) = \frac{1}{2}\sum_{ij}\left|m_{ij} - b_{ij}\right|$. The following constraints should be satisfied for b_{ij}: $b_{ik} - \sum_{j=1}^{k-1} b_{j\,j+1} = 0$; (i=1,…, n-2; k=i+2,…,n and $1-n \le b_{ij} \le n-1$; b_{ij} integer).