ELSEVIER

European Journal of Operational Research 123 (2000) 593–605

EUROPEAN JOURNAL OF OPERATIONAL RESEARCH

www.elsevier.com/locate/dsw

Theory and Methodology

Comparing the efficacy of ranking methods for multiple round-robin tournaments

D. Mendonça *, M. Raghavachari

Department of Decision Sciences and Engineering Systems, Rensselaer Polytechnic Institute, 110 8th Street, Troy, NY 12180, USA

Received 1 July 1998; accepted 1 January 1999

Abstract

A methodology is developed for comparing the efficacy of a number of multiple round-robin tournament ranking methods with respect to their ability to replicate the true rank order of players' strengths. Four new ranking methods are introduced in this paper. Different probability models for pairwise contests between tournament participants are considered. We consider the case in which there are multiple tournaments but not all players compete in all tournaments. Guidelines are provided for selecting a ranking method given some foreknowledge of the form of the distributions of the participants' strengths. © 2000 Elsevier Science B.V. All rights reserved.

Keywords: Modeling; Simulation; Tournaments; Ranking

1. Introduction

The purpose of this paper is to evaluate various methods for ranking a set of players who compete in round-robin tournaments when not all tournaments involve every player. Additionally, four new ranking methods for such multiple incomplete tournaments are proposed in this paper. A simulation methodology is developed in which two different models of player strength are considered. By beginning with models of player strength, the true strengths of players are therefore known, so that complete rankings of these strengths are obtainable. A set of round-robin tournaments is then induced and a ranking obtained via each ranking method. The four new methods and one from the tournament ranking literature are then compared for a sample problem with respect to their properties and their ability to reflect the true ranking.

The ranking methods evaluated here may be used to rank objects such as players in round-robin tournaments when not all objects may be compared along all criteria. Potential applications may be found in a number of areas. In evaluations of a number of related consumer products, the criteria of taste, color and smell may not be present in all the objects but it may still be desirable to form an overall ranking (Cook et al., 1996). In committee decision making, such as for project selection, not all projects may be comparable along all criteria;

* Corresponding author. Tel.: +1-518-276-2773; fax: +1-518-276-8227.

PII: S 0 3 7 7 - 2 2 1 7 (9 9) 0 0 1 1 0 - 1

however, it is still necessary to rank the projects overall (Cook et al., 1996). In conducting company-wide performance evaluations of employees (David, 1988), some performance criteria (e.g., total sales) may not be applicable to all employees. Certain sporting competitions may also exhibit this structure.

Our research methodology is novel for two reasons: first, ranking efficacy is evaluated with respect to the known (but in practice unobservable) true ranking; second, performance of the methods is evaluated under different mixes of player strengths. The evaluation methodology developed in this research can help guide the selection of a ranking method in pairwise comparison experiments having round-robin structure, even when not all objects can be judged on all criteria. For the sake of convenience, these objects will be referred to as players in the tournament context.

The proposed evaluation methodology involves inducing a schedule of round-robin tournaments under different models of player strength, ranking the players via various methods and testing the efficacy of the methods in reproducing the true ranking. In the following section, two probability models on pairwise comparisons are briefly reviewed. Next, in Section 3, we discuss the five ranking methods. In Section 4, the simulation methodology is presented and the results summarized in Section 5. Finally we discuss the results and offer concluding remarks in Section 6.

2. Background

2.1. Probability models on pairwise comparisons

A model of pairwise comparisons attempts to explain the capability of one player to defeat another in a single contest. Assume that the strength of each player i, $i = 1, \ldots, n$, is modeled by a random variable X_i having mean V_i. The ranking of the mean strengths of the players can then be taken as the true ranking.

Two well-known models for player strength are: (1) that of Bradley and Terry (1952), Bradley (1953), Mallows (1957) [also attributed to Zermelo (David, 1988)] and (2) that of Thurstone and Mosteller (David, 1988). If π_{ij} is the probability that player i defeats player j, then under the first model (denoted BTM),

$$\pi_{ij} = V_i/(V_i + V_j). \tag{1}$$

Several distributions of the X_is may yield BTM and it may be verified that (1) follows if each X_i is exponentially distributed with mean V_i (David, 1988). This observation is used in our simulation methodology as a means of simulating pairwise comparisons under BTM. Under the Mosteller–Thurstone model (denoted MT), each V_i is the mean of normally distributed random variable X_i with variance σ^2. Therefore,

$$\pi_{ij} = \Phi\left(\frac{V_i - V_j}{\sigma\sqrt{2}}\right), \tag{2}$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. The simulation methodology of this paper assumes common variance of the distributions of strengths of players. Note, however, that the methodology can accommodate the case of unequal but known variances. The difficult case of unequal and unknown variances is a matter for further study.

2.2. Ranking methods for single tournaments

In a single round-robin tournament, each one of n players plays the remaining $(n - 1)$ players exactly once. Results of a tournament between n players can be summarized in an $n \times n$ incidence (i.e., tournament) matrix **A** whose (i, j)th element is 1 if player i defeats player j and is 0 otherwise (it is assumed that no draws are allowed). Methods for determining rankings in tournaments are an outgrowth of work on evaluating the results of pairwise comparisons. Surveys are given by David (1988) for pairwise comparisons, Moon (1968) for tournaments generally, Vol'skii (1988) for ranking methods and Marden (1995) for modeling rank data.

A method of ranking participants based on row sums of the incidence matrix was first proposed by Kendall (1955). Player i's relative strength, R_i, is determined as follows:

$$R_i = \frac{a_i}{(n-1)}, \tag{3}$$

where a_i is the sum of the elements in the ith row of **A**; that is, the number of pairwise comparisons won by player i. Kendall's method is simple to use and has considerable intuitive appeal.

Wei (1952) as cited in Cook and Kress (1992) and Kendall (1955) proposed applying a theorem of Frobenius (Moon, 1968) to the incidence matrix **A** to generate **R**, the vector of players' relative strengths, as follows:

$$\mathbf{R} = \lim_{\ell \to \infty} \left(\frac{\mathbf{A}}{\lambda}\right)^{\ell} \times \boldsymbol{e}, \tag{4}$$

where $\boldsymbol{e}$ is an $n \times 1$ vector of ones and λ is the unique positive characteristic root of **A** with largest absolute value.

In the following section, extensions to the multiple tournament ranking problem are proposed for the methods given by Eqs. (3) and (4); a method given by Cook et al. (1996) is described and a modification of it introduced. After defining these methods, we discuss the ways in which they are to be compared.

3. Ranking methods for multiple tournaments

In multiple round-robins, players compete in a number of tournaments. Ranking methods for many multiple complete tournament structures are reviewed by McGarry and Schutz (1997). We are concerned with multiple incomplete tournaments held among n players. Some of the T tournaments are incomplete in that they are contested by $n_t \leqslant n$ players, thus yielding $n_t(n_t - 1)/2$ outcomes for tournament t. A ranking method for multiple incomplete tournaments allows the comparison of n players even when not all players participate in all tournaments.

In order to discuss ranking methods for multiple tournaments, some additional notation (adapted from Cook et al., 1996) must be introduced. Let K_{i_0} be the set of tournaments in which player i_0 competes and let $a^t_{i_0}(m)$ be the row sum for player i_0 in the mth power of the incidence matrix of tournament t. Let R_{i_0} be the score assigned to player i_0 by a ranking method. The notation i_0 is necessary since, for example, the second player in tournament may not be "Player 2". Finally, let $|B|$ denote the cardinality of any set B.

3.1. Method PT

Cook et al. (1996) present a number of linear programming formulations of the ranking problem for players in multiple incomplete tournaments. The one which is considered here is labeled PT (Cook et al., 1996, pp. 874–875). Note that in implementing PT, individual tournament weights (w_t) were held equal for $t = 1, \ldots, T$.

3.2. Method PT modified

PT is extended in this paper by recognizing that the number of Hamiltonian paths in a tournament t is $M_t = (n_t - 1)$ (Moon, 1968). Therefore, changes in PT need only be made to the objective function

$$R_{i_0} = \max_{\alpha^t(m)} \frac{\sum_{t \in K_{i_0}} \sum_{m=1}^{M_t} \alpha^t(m) a^t_{i0}(m)}{\left|K_{i_0}\right|} \tag{5}$$

and to the following constraints:

$$\sum_{t \in K_i} \sum_{m=1}^{M_t} \alpha^t(m) a^t_{i.}(m) \leqslant |K_{i.}|, \quad i = 1, \ldots, n, \tag{6}$$

$$\sum_{m=1}^{M_t} \alpha^t(m) a^t_{i.}(m) \leqslant 1, \quad i = 1, \ldots, n, \ t \in K_{i.}, \tag{7}$$

$$\alpha^t(m) - \alpha^t(m+1) \geqslant \varepsilon, \quad m = 1, \ldots, (M_t - 1), \quad t = 1, \ldots, T, \tag{8}$$

$$\alpha^t(M_t) \geqslant \varepsilon, \quad t = 1, \ldots, T, \tag{9}$$

where $a^t_{i.}(m)$ is the row sum for player i in the mth power of the incidence matrix for tournament t. This method will be referred to as PTM, for PT Modified.

3.3. Method Kendall modified

In this paper, Kendall's method for single tournaments is modified for the multiple tournament case as follows:

$$R_{i_0} = \sum_{t \in K_{i_0}} \frac{a^t_{i_0}(1)}{|K_{i_0}|}. \quad (10)$$

Eq. (10) therefore represents the average winning percentage of player i_0 taken across the tournaments in which that player competes. We will refer to this method as KM for Kendall modified.

3.4. Method eigenvector 1

The first eigenvector-based method is similar in construction to KM:

$$R_{i_0} = \sqrt{\frac{\sum_{t \in K_{i_0}} u^2_{i_0}(t)}{|K_{i_0}|}} \quad (11)$$

with which is associated a vector

$$\boldsymbol{u}^*(\boldsymbol{t}) = \lim_{\ell \to \infty} \left(\frac{\mathbf{A}^t}{\lambda_t} \right)^\ell \times \boldsymbol{e}, \quad (12)$$

the non-normalized score vector for tournament $t \in K_{i_0}$, for which λ_t is the is the unique positive characteristic root with largest absolute value for the incidence matrix of tournament t. In Eq. (11), $u^2_{i_0}(t)$ is the square of the element of the normalized form of $\boldsymbol{u}^*(\boldsymbol{t})$ corresponding to player i_0. The score is therefore an average of the relative strengths across all tournaments in which i_0 competes. We will refer to this method as Eig1.

3.5. Method eigenvector 2

The second eigenvector method is based on the distance from a vector whose elements are a player's relative strengths (i.e., the $|K_{i_0}|$ values of $u^2_{i_0}(t)$ where $t \in K_{i_0}$) to some optimal vector. Consider a schedule of three complete tournaments (i.e., $n_t = n$ for all t). Each of the three values of $u^2_{i_0}(t)$ can at most be unity. So, in T-dimensional space, a vector of optimal scores can be represented as $\boldsymbol{e}$, the $n \times 1$ vector of ones. Let the vector

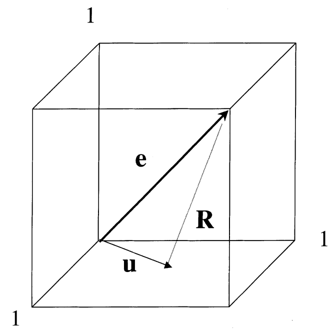

Fig. 1. Relative strength of one player ($T = 3$).

of player i_0's relative strengths be represented by $\boldsymbol{u}_{i_0}$ and let $\boldsymbol{e}_{i_0}$ be a unit vector of dimensionality $|K_{i_0}|$. The situation is shown in Fig. 1.

We therefore propose that a player's performance be reckoned according to distance from $\boldsymbol{u}_{i_0}$ to $\boldsymbol{e}_{i_0}$. A player's relative strength may then be based on the head-to-head distance:

$$R_{i_0} = \frac{1}{\|\boldsymbol{u}_{i_0} - \boldsymbol{e}_{i_0}\|}. \quad (13)$$

This method will be referred to as Eig2.

4. Simulation methodology

We now describe a simulation methodology for testing the efficacy of the multiple tournament ranking methods described in the previous section. The methodology is summarized as follows. Beginning with known mean strengths of players, a schedule of multiple incomplete tournaments is simulated under two different models of player strength. The tournament outcomes are supplied to the various ranking methods to produce corresponding estimates of the true ranking. Finally, the efficacy of the ranking methods in reproducing the true ranking is tested. The remainder of this section details the methodology along with an example problem.

4.1. Assign mean strengths to players

The following six different profiles of mean strength, representing six different configurations

Table 1
Mean player strengths under various profiles

Profile	Player					
	1	2	3	4	5	6
1	1	1	1	1	1	1
2	1	2	3	4	5	6
3	1	3	5	7	9	11
4	7	8	9	10	11	12
5	7	9	11	13	15	17
6	7	10	13	16	19	22

of players, are used in simulating a schedule of incomplete tournaments (see Table 1).

Note that these profiles allow for a very full investigation of the relative performance of the methods. Both relative and absolute strengths are varied. For example, mean strengths of adjacent players in profiles two and four differ by one unit but have different absolute sizes. Profiles 2 and 3 have greater spread between adjacent mean strengths but the strength of the weakest player is unity in each profile.

4.2. Determine schedule of tournaments

We consider the schedule for $n = 6$ players and $t = 4$ tournaments given by Cook et al. (1996) (shaded areas indicate participation) (see Table 2).

4.3. Determine match outcomes

The $n_t(n_t - 1)/2$ match outcomes for a tournament t are simulated under BTM and MT. Recall that under BTM, an exponential random variable X_i having mean V_i serves as the model of player i's strength and that under MT, X_i is a normal random variable having mean V_i and variance σ^2. The variance σ^2 for all profiles under MT is here held constant at 3, which we regard as reasonably appropriate for the range of player strengths. A particular match outcome is determined via simulation as follows: if $x_i \geqslant x_j$ for a tournament t, then the (i, j)th element of $\mathbf{A}^t$ is one; otherwise, it is zero. 1000 simulations were performed under each of the two models.

Table 2
Schedule of tournaments

	Player					
Tournament	1	2	3	4	5	6
1	■	■	■	■		
2	■	■		■	■	
3	■	■	■		■	■
4	■			■	■	■

4.4. Apply and test ranking methods

Each one of the various ranking methods is then applied to the results of the tournament schedule, resulting for each simulation in a vector of players' raw scores for each method. The players are then ranked and the ranking compared to the true ranking as determined by the mean strength. Let y_{ki_0} be the rank of player i_0 according to method k and let s_{i_0} be the true rank of player i_0. The statistic of interest is the well-known Spearman's rank correlation coefficient (Sachs, 1984), referred to here as Spearman's r. For a method k, the statistic for the jth simulation is computed as

$$r_{jk} = 1 - \frac{6\sum_{i_0=1}^{n}\left(y_{jki_0} - s_{i_0}\right)^2}{n^2(n-1)}. \tag{14}$$

For a given profile of players a one-way analysis of variance (ANOVA) is performed to test equality of the average correlation coefficients among all five methods (i.e., PT, PTM, KM, Eig1, Eig2). If ANOVA determines that any means are unequal, Tukey's multiple comparisons procedure (Sachs, 1984) is applied to cluster the methods into groups having statistically equal means.

4.5. Note on implementation

Tournament outcomes were generated using built-in number generation in AMPL®. The rankings according to methods PT and PTM were obtained using AMPL® in conjunction with the solver CPLEX®. The rankings according to methods KM, Eig1 and Eig2 were obtained in MATLAB®. All statistical tests were performed with SAS®.

5. Results

The above schedule was run one thousand times for each of the two models of player strength. Results of the simulations are summarized in this section and discussed in Section 6.

5.1. Bradley–Terry/Mallows (BTM)

The averages of Spearman's correlation coefficient for each method under each profile are given in Table 3.

To illustrate the table, in profile two the average correlation coefficient between the ranking obtained via method KM and the true ranking is 0.66.

ANOVA (not shown here) reveals that for profile five the average correlation coefficients are equal for all methods. For each of the other profiles, ANOVA rejects the hypothesis that the average correlation coefficients are equal ($p \leqslant 0.0001$).

In a further analysis of the correlation coefficients, application of Tukey's procedure to profiles 1, 2, 3, 4 and 6 yields simultaneous 95% confidence intervals for all pairwise differences in the means. The resulting groups are indicated by parentheses in the following table. Note that $A > B$ indicates that the average correlation coefficient for method A was significantly greater than that for method B at $\alpha = 0.05$ (see Table 4).

5.2. Box plots of Spearman's r under BTM

The shape and spread of the empirical distributions of the correlation coefficients for each of the six methods in profiles one to six are summarized in the following box plots. In the plots, a circle indicates an outlier; a star indicates an extreme outlier (see Figs. 2–7). A comparison of the box plots is given in Section 6.

A study of Pearson's product–moment correlation between pairs of methods across profiles two through six reveals the following. Methods PT and PTM are highly correlated (approximately 0.98), KM is correlated with PT and PTM (approximately 0.82), Eig1 is correlated with PT, PTM and KM at about 0.68, Eig2 is negatively correlated (approximately −0.08) with all other methods. For profile one, all methods have correlation of approximately zero with one another, except for PT and PTM, which have correlation equal to 0.74. Thus it can be seen that the degree of agreement between the methods themselves is highly variable.

Table 3
Average correlation coefficient for various methods

Profile	Method				
	PT	PTM	KM	Eig1	Eig2
1	0.50	0.50	0.50	0.50	0.50
2	0.63	0.63	0.66	0.56	0.49
3	0.69	0.69	0.71	0.61	0.50
4	0.27	0.28	0.29	0.35	0.44
5	0.43	0.42	0.45	0.44	0.43
6	0.51	0.51	0.53	0.48	0.46

Table 4
Efficacy of various ranking methods under BTM

Profile	Result
1	KM > Eig1 > (Eig2, PT, PTM)
2	(KM, PT, PTM) > Eig1 > Eig2
3	(KM, PT, PTM) > Eig1 > Eig2
4	Eig2 > Eig1 > (KM, PT, PTM)
6	(KM, PT, PTM) > Eig2 but neither significantly different from Eig1

5.3. Mosteller–Thurstone (MT)

The average correlation coefficients for the various methods are shown in Table 5. The entries in Table 5 for the MT model are interpreted in the same way as those in Table 3 for the BTM model.

ANOVA reveals that at least two average correlation coefficients within every profile are unequal ($p \leqslant 0.0001$). Tukey's procedure ($\alpha = 0.05$) yields the results in Table 6.

5.4. Box plots of Spearman's r under MT

As is the case under BTM, the degree of agreement within each profile among pairs of methods is

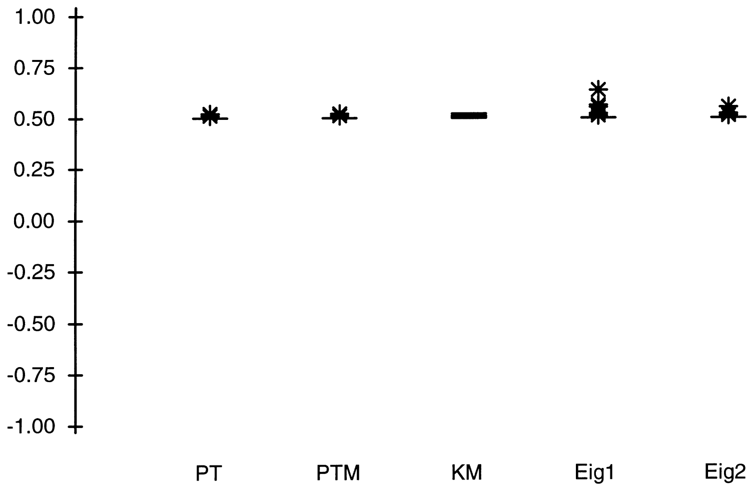

Fig. 2. Box plot of Spearman's r for profile one under BTM, all methods.

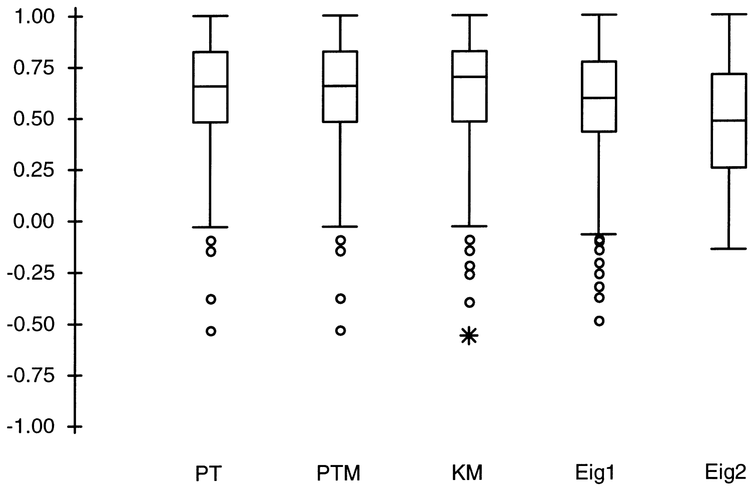

Fig. 3. Box plot of Spearman's r for profile two under BTM, all methods.

Table 5
Average correlation coefficient for various methods

Profile	Method				
	PT	PTM	KM	Eig1	Eig2
1	0.50	0.50	0.50	0.50	0.50
2	0.84	0.84	0.87	0.75	0.46
3	0.94	0.93	0.94	0.85	0.42
4	0.84	0.84	0.86	0.76	0.46
5	0.93	0.93	0.94	0.85	0.43
6	0.97	0.96	0.97	0.88	0.42

highly variable as measured by Pearson's product–moment correlation. For profiles two to six, PT and PTM are highly correlated (approximately 0.98), the correlation of KM with PT and PTM is approximately between 0.70 and 0.85, Eig1 is correlated with PT, PTM and KM at approximately between 0.40 and 0.60 and Eig2 is negatively correlated with all other methods (approximately between −0.08 and −0.46). For profile one, all correlations are less than 0.13, except PT and PTM,

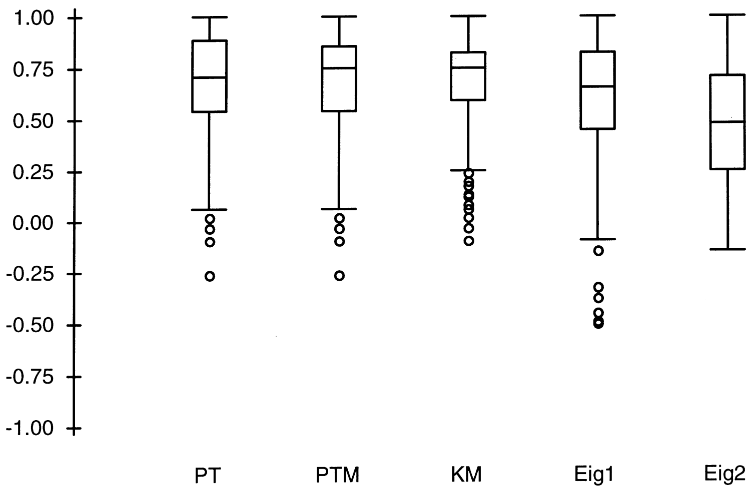

Fig. 4. Box plot of Spearman's r for profile three under BTM, all methods.

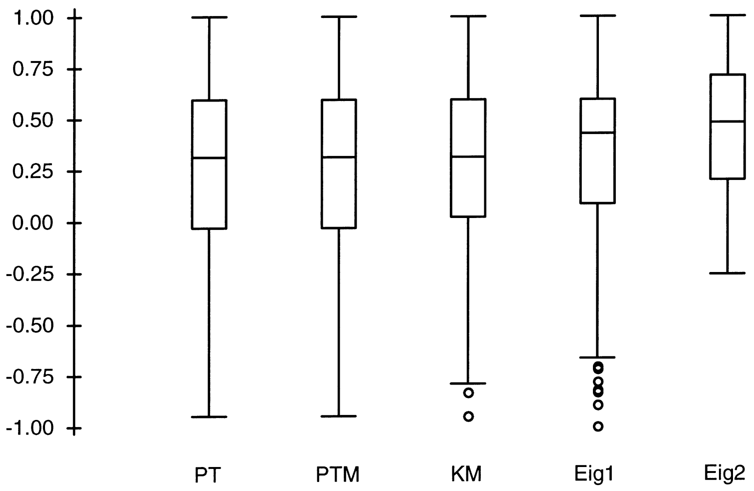

Fig. 5. Box plot of Spearman's r for profile four under BTM, all methods.

Table 6
Efficacy of various ranking methods under MT

Profile	Result
1	KM > Eig1 > (PT, Eig2) but neither the second nor third group is significantly different from PTM
2	KM > (PT, PTM) > Eig1 > Eig2
3	(KM, PT, PTM) > Eig1 > Eig2
4	KM > (PT, PTM) > Eig1 > Eig2
5	(KM, PT, PTM) > Eig1 > Eig2
6	(KM, PT, PTM) > Eig1 > Eig2

which have correlation of 0.70, and Eig1 and Eig2, which have correlation 0.40 (see Figs. 8–13).

6. Discussion and conclusions

6.1. Simulation results and conclusions

The methodology developed for testing the efficacy of ranking methods allows the decision

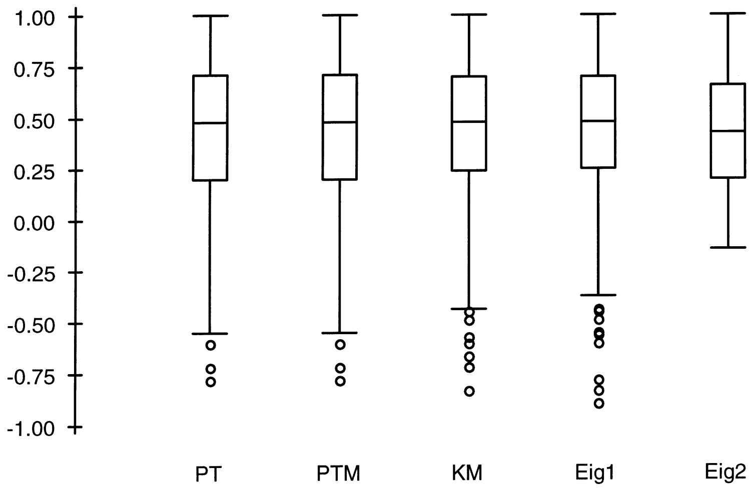

Fig. 6. Box plot of Spearman's r for profile five under BTM, all methods.

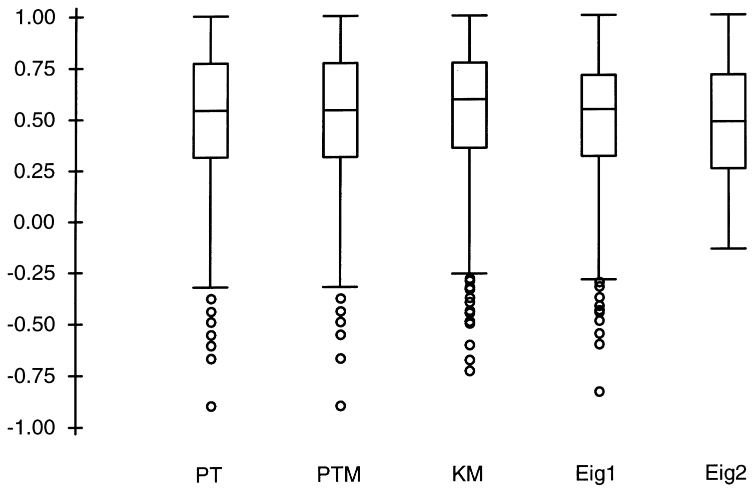

Fig. 7. Box plot of Spearman's r for profile six under BTM, all methods.

maker to reason forward from assumptions about the distribution of players' strengths in order to assess which method or methods are most likely to reflect the true ranking of players. By considering various profiles of players' strengths in a given schedule, a decision maker may explore the ranking methods' efficacy for varying degrees of differences in player strengths, as well as in the magnitude of those strengths.

6.1.1. Bradley–Terry/Mallows (BTM)

Based on the tests discussed in Section 5, the average correlation for the ranking methods is no more than 71%, as shown in Table 3, and is often quite lower. The methods perform differently within each profile, as shown in Table 4, although KM is usually in the top group. Methods PT, PTM, KM and Eig1 have greater skew and, with the exception of profile three, greater range rela-

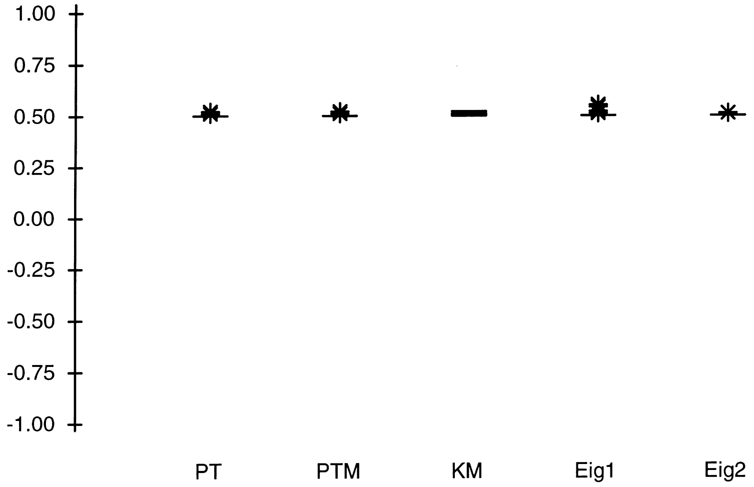

Fig. 8. Box plot of Spearman's r for profile one under MT, all methods.

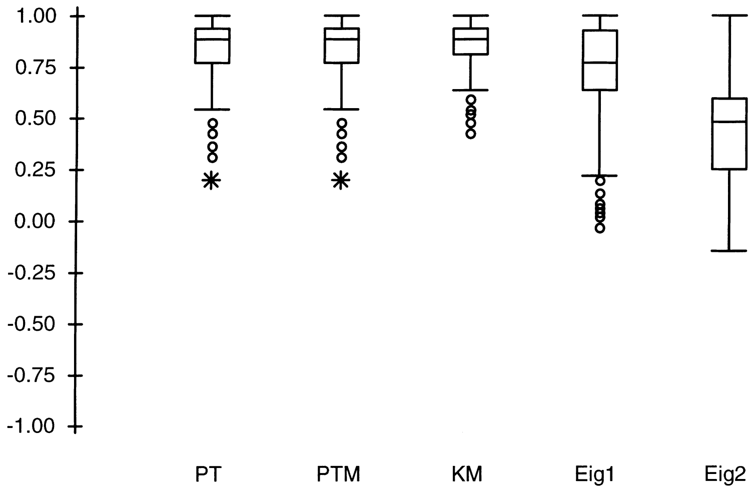

Fig. 9. Box plot of Spearman's r for profile two under MT, all methods.

tive to Eig2. The obverse of the method's symmetry and stability is lower average correlation: for all but profile four, Eig2 is a member of the lowest group. Given the attractive properties of the method Eig2 under BTM, its further refinement is an area of future work.

Looking at individual methods across profiles, the shape of the distribution of the correlation coefficient for method Eig2 is generally unchanged and therefore not as sensitive as the other methods to changes in the profiles of players' strengths. An increase in the gap between individual mean player strengths produces higher average correlations in PT, PTM, KM and Eig1, though the distributions remain asymmetric.

6.1.2. Mosteller–Thurstone (MT)

According to Table 5, the average correlation coefficients for PT, PTM and KM are, with the exception of profile one, greater than 84%.

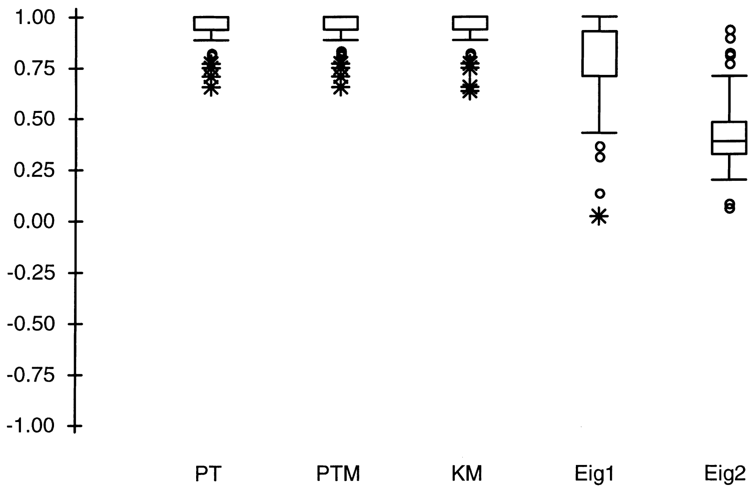

Fig. 10. Box plot of Spearman's r for profile three under MT, all methods.

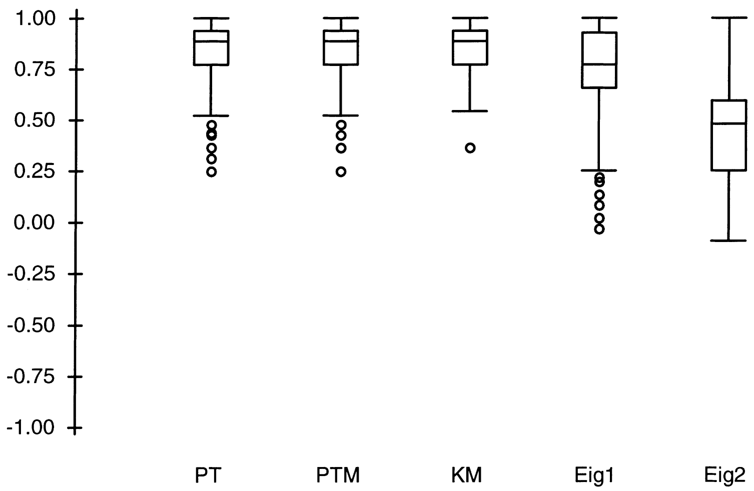

Fig. 11. Box plot of Spearman's r for profile four under MT, all methods.

Tukey's procedure places Eig1 and Eig2 in the last and second-to-last groups respectively for all profiles. The variabilities in the average correlation coefficients for PT, PTM and KM are low relative to Eig1 and Eig2. Moreover, as shown in Table 6, PT, PTM and KM are placed in the first or second groups for all profiles by Tukey's procedure.

In contrast to simulations under BTM, those under MT show high average correlation for methods PT, PTM and KM for all profiles. Under MT, then, the rankings obtained via these three methods are expected to agree often for profiles similar to those considered here.

Under BTM, the pairwise probabilities for all players are more nearly equal than under MT.

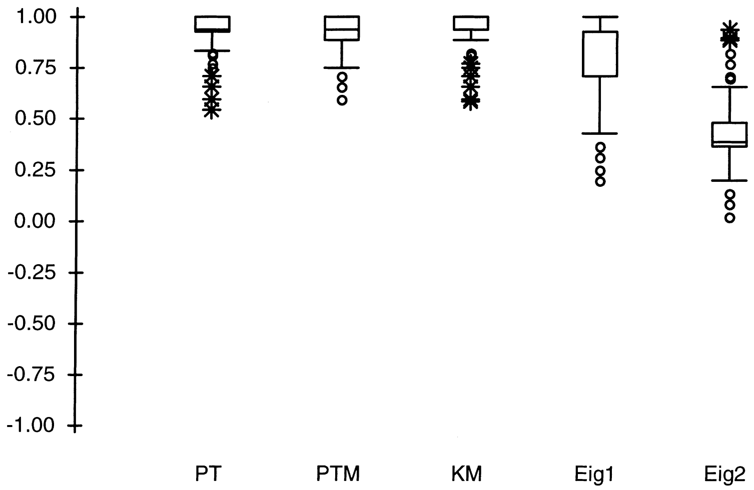

Fig. 12. Box plot of Spearman's *r* for profile five under MT, all methods.

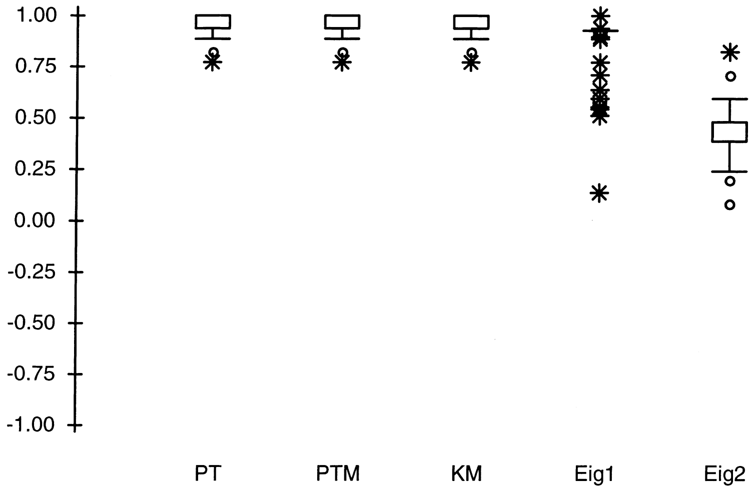

Fig. 13. Box plot of Spearman's *r* for profile six under MT, all methods.

Note, for example, that performance of the methods under BTM is nearly always lower for profiles four through six compared to profiles two and three. Under MT, performance of the models is nearly identical across these two sets of profiles. Methods generally perform better under MT than under BTM.

The three top-performing measures (KM, PT and PTM) share a common property in that they take degree of participation into account and place reduced (PT and PTM) or no (KM) weight on *m*th generation wins. To a lesser extent, method Eig1 accounts for degree of participation, but it seems that the effect of including *m*th generation wins has a deteriorating effect on its performance. Similarly, Eig2, which takes no account of degree of participation, performs poorly as the ranking it provides is based entirely on *m*th generation wins; however, it should be noted that the performance of method Eig2 is relatively unaffected by the underlying model. With Eig2, even a very strong showing by player 6 (the true strongest player) may not be as well-rewarded as mediocre showings by those players participating in three or four tournaments. A repositioning of the optimum vector, perhaps to account for the tournament difficulty or number of players per tournament, may improve the efficacy of Eig2.

The simulation methodology has been implemented under a particular set of conditions comprised of certain degrees of participation by players, a certain set of profiles and a certain schedule. Under such conditions, KM may be the most efficacious method as evaluated by Spearman's correlation coefficient. Additionally, KM's ease of implementation may make it easier to explain and justify than methods which incorporate mth generation wins. While methods PT and PTM are strong performers, particularly under MT, neither is ever in a group which outperforms KM. When the BTM model governs the pairwise comparisons and if the cost of making an error increases with the degree of error (again, as measured by Spearman's r), method Eig2 ought to be considered. While its average correlation is generally lower than that of the other methods, its errors are generally not so severe.

6.2. Possible extensions

A number of extensions to the current implementation are possible. The sensitivity of the methods to changes in the profiles may be investigated more fully. For example, a larger number of profiles may be randomly generated according to different means and variances of the strengths' distributions. It would then be possible to draw conclusions about profiles which are generated with unequal variances and constant means, and vice versa. Finally, it may be desirable to investigate the impact of different tournament schedules on the efficacy of the methods.

Acknowledgements

The authors thank an anonymous referee for helpful comments and suggestions which improved the quality of the presentation.

References

Bradley, R.A., 1953. Some statistical methods in taste testing and quality evaluation. Biometrics 9, 22–38.

Bradley, R.A., Terry, M.E., 1952. Rank analysis of incomplete block designs: I. The method of paired comparisons. Biometrika 39, 324–345.

Cook, W.D., Doyle, J., Green, R., Kress, M., 1996. Ranking players in multiple tournaments. Computers and Operations Research 23/9, 869–880.

Cook, W.D., Kress, M., 1992. Ordinal Information and Preference Structures. Prentice Hall, Englewood Cliffs, NJ.

David, H.A., 1988. The Method of Paired Comparisons, 2nd ed., vol. 41. Oxford University Press, New York.

Kendall, M.G., 1955. Further contributions to the theory of paired comparisons. Biometrics 11, 43–62.

Mallows, C.L., 1957. Non-null ranking models: I. Biometrika 44, 114–130.

Marden, J.I., 1995. Analyzing and Modeling Rank Data, vol. 64. Chapman & Hall, New York.

McGarry, T., Schutz, R.W., 1997. Efficacy of traditional sport tournament structures. Journal of the Operational Research Society 48, 65–74.

Moon, J.W., 1968. Topics on Tournaments. Holt, Rinehart and Winston, New York.

Sachs, L., 1984. Applied Statistics (Zenon Reynarowych, Trans.), 2nd ed. Springer, New York.

Vol'skii, V.I., 1988. Choice of best alternatives on directed graphs and tournaments. Avtomatika i Telemekhanika 3, 3–17.

Wei, T.H., 1952. The Algebraic Foundations of Ranking Theory. Cambridge University, Cambridge, UK.