3-2008

Testing intergroup concordance in ranking experiments with two groups of judges

Dawn J. DEKLE
Singapore Management University

LEUNG, Denis H. Y.
Singapore Management University, denisleung@smu.edu.sg

Min ZHU
CSIRO

DOI: https://doi.org/10.1037/1082-989X.13.1.58

Published in Psychological Methods, March 2008, 13 (1) pp. 58-71.
http://doi.org/10.1037/1082-989X.13.1.58

Testing Intergroup Concordance in Ranking Experiments With Two Groups of Judges

Dawn J. Dekle and Denis H. Y. Leung
Singapore Management University

Min Zhu
Commonwealth Scientific and Industrial Research Organisation

Across many areas of psychology, concordance is commonly used to measure the (intragroup) agreement in ranking a number of items by a group of judges. Sometimes, however, the judges come from multiple groups, and in those situations, the interest is to measure the concordance between groups, under the assumption that there is some within-group concordance. In this investigation, existing methods are compared under a variety of scenarios. Permutation theory is used to calculate the error rates and the power of the methods. Missing data situations are also studied. The results indicate that the performance of the methods depend on (a) the number of items to be ranked, (b) the level of within-group agreement, and (c) the level of between-group agreement. Overall, using the actual ranks of the items gives better results than using the pairwise comparison of rankings. Missing data lead to loss in statistical power, and in some cases, the loss is substantial. The degree of power loss depends on the missing mechanism and the method of imputing the missing data, among other factors.

Keywords: concordance, intergroup, Kendall's *W*, missing data, ranking experiment

Ranking is commonly used in empirical psychological research to measure the preference of an individual who is presented with a set of alternatives (e.g., Bonner, 2004; Castel, Miró, & Rull, 2005; Elstein, Chapman, & Knight, 2005; Fisher, Macrosson, & Yusuff, 1996; Marlowe, Schneider, & Nelson, 1996; Miró, Huguet, & Nieto, 2005; Stewart & Stewart, 1996; Swanson, Wigal, & Udrea, 1998; Wanschura & Dawson, 1974). In a ranking experiment, a number of participants (often called *judges* in the literature) are presented with a set of K alternatives, and each participant is asked to rank the alternatives (sometimes called *items*) from the most preferred to the least preferred.

Sometimes, the judges come from G different groups (Lohmann, Delius, Hollard, & Friesel, 1988; McKnight & Hills, 1999; Pope & Scott, 2003; Rule, Bisanz, & Kohn, 1985; Vidmar & Cernigoj, 2004). Then the interest may be to determine the degree of agreement or concordance between groups in ranking the K alternatives. A number of studies have appeared in the literature that addresses this problem. For $G = 2$, Linhart (1960) suggested calculating a Kendall's W (Kendall & Smith, 1939; Kendall & Gibbons, 1990) in each of the two groups of judges and then comparing the difference. However, this method is a test for within-group rather than between-groups concordance. In an alternative solution, Hays (1960) used Kendall's τ (Kendall & Smith, 1939) for the rankings between pairs of judges as an overall measure of agreement in the rankings between judges. Following that line of reasoning, in a series of articles, Schucany and colleagues (Beckett & Schucany, 1979; Li & Schucany, 1975; Schucany & Beckett, 1976; Schucany & Frawley, 1973) calculated the average of Spearman's r_s (Spearman, 1904) between pairs of judges, one drawn from either group. Works related to Schucany and colleagues can also be found in Legendre (2005), Lyerly (1952), and Page (1963), whose interests are primarily in measuring concordance between a single judge and a group of other judges. Hollander and Sethuraman (1978) used a statistic based on the Mahalanobis distance (Mardia, Kent, & Bibby, 1979, p. 16) of the average rankings between groups. Finally, Kraemer (1981) defined intergroup concordance as the ratio of the intergroup Kendall's W on the basis of the mean ranks of each group to the average of the within-group Kendall's Ws. Feigin and Alvo (1986) proposed a statistic similar to Kraemer's statistic by using the ratio of the average diversity within groups to the diversity between groups. The diversity measure can be based on Kendall's τ, Spearman's r_s, or Spearman's footrule (Diaconis & Graham, 1977).

Despite the large amount of existing research on measur-

Dawn J. Dekle and Denis H. Y. Leung, School of Economics, Singapore Management University, Singapore; Min Zhu, Commonwealth Scientific and Industrial Research Organisation, Marine Laboratories, Cleveland, Queensland, Australia.

Dawn J. Dekle and Denis H. Y. Leung were partially funded by a grant from the Research Center, Singapore Management University.

Correspondence concerning this article should be addressed to Denis H. Y. Leung, Singapore Management University, 90 Stamford Road, Singapore 178903. E-mail: denisleung@smu.edu.sg

ing intergroup concordance, no formal comparison of the different methods has been made. Evaluation of existing work is complicated by the absence of a commonly agreed on set of hypotheses (see discussions in Beckett & Schucany, 1979; Hays, 1960; Hollander & Sethuraman, 1978; Kraemer, 1981; Li & Schucany, 1975; Schucany & Frawley, 1973). Li and Schucany (1975, p. 419) and Kraemer (1981, p. 645) discussed the relationships between their methods and earlier methods but did not carry out formal comparisons. Furthermore, in most existing works, the large sample properties of their methods were studied. In most applications, however, the sample size (number of judges $\times$ number of items) is unlikely to be large, leaving the question of whether large sample results could be extrapolated to those situations. Finally, missing data are a common problem in empirical research (Little & Rubin, 2002). However, only a few articles deal with the treatment of missing data in calculating rank concordance (Alvo & Cabilio, 1995; Schucany & Beckett, 1976; Stephens, Claypool, & Buchalter, 1977, 1978; Yu, Lam, & Alvo, 2002). The methods of both Schucany and Beckett (1976) and Stephens et al. (1977, 1978) assumed the special case that every judge in each group leaves the same number of items unranked. Alvo and Cabilio (1995) suggested an imputation method for data that are missing completely at random (Rubin, 1987). Yu et al. (2002) extended the method of Alvo and Cabilio (1995) to situations where there may be ties in the data.

Our primary aim in this article is to carry out a comparison of the existing methods using a simulation study. Instead of using large sample theory, we use nonparametric evaluations, via permutation theory, to determine the sampling distributions of the test statistics. The reason for using permutation theory is twofold. First, permutation theory is more robust than large sample theory in small and moderate sample situations. Second, large sample theory is affected by missing data in the sense that adjustments in the test statistic are required for the theory to remain valid for drawing inferences. These adjustments are often difficult to derive analytically. However, no adjustments are needed to derive the permutation distributions. An additional hurdle to be overcome in this study is in generating data for the simulations. True rank data under specific levels of concordance are practically impossible to generate. However, this difficulty can be overcome by assuming that the ranking data follow a latent rating model (Thurstone, 1927, 1931). In a latent rating model, ranks are induced by latent ratings given to different items. In this study, a latent multivariate normal rating model is used to generate the ratings, and specific levels of concordance can be induced using different parameters in the multivariate normal model.

Existing Methods

Preliminaries

In this study, we consider the case where there are two groups of J judges each. The generalization to situations with more than two groups of judges and unequal numbers of judges per group will be deferred to the Discussion section. It is assumed that each judge is asked to rank K items. For each item, the judge is supposed to return a rank from 1 to K, indicating the judge's preference for the item (1 = most preferred and K = least preferred). Items could receive the same rank (in which case the mean rank could be used to assign ranks; see, e.g., Li & Schucany, 1975), and it is possible that a judge may leave some items unranked, an issue that we explore later. For ease of illustration, it is assumed that there are no missing data or tied rankings. Let R_{gjk} be the ranking of the kth item by the jth judge from the gth group, $k = 1, \ldots, K$; $j = 1, \ldots, J$; $g = 1, 2$. Furthermore, let $R_{g\cdot k}$, $\bar{R}_{g\cdot k}$, and $\bar{R}_{\cdot\cdot k}$ be, respectively, the sum and the average ranking of item k by the J judges in group g and the overall average ranking of item k by the $2J$ judges in both groups. Then Spearman's correlation, r_s, of the rankings between two typical judges j and j', when both are from the first group, both are from the second group, and one is from each of the two groups are defined, respectively, as

$$\hat{r}_{s,g(jj')} = 1 - \frac{6}{K^3 - K}\sum_{k=1}^{K}(R_{gjk} - R_{gj'k})^2, \qquad g = 1, 2,$$

$$\hat{r}_{s,1(j)2(j')} = 1 - \frac{6}{K^3 - K}\sum_{k=1}^{K}(R_{1jk} - R_{2j'k})^2.$$

Similarly, the corresponding Kendall's τ (see, e.g., Alvo & Cabilio, 1995; Hays, 1960; Yu et al., 2002)[1] can be defined as

$$\hat{\tau}_{g(jj')} = \frac{2}{K^2 - K}\sum_{1\le k<k'\le K}\operatorname{sgn}(R_{gjk} - R_{gjk'})\operatorname{sgn}(R_{gj'k} - R_{gj'k'}), \qquad g = 1, 2,$$

$$\hat{\tau}_{1(j)2(j')} = \frac{2}{K^2 - K}\sum_{1\le k<k'\le K}\operatorname{sgn}(R_{1jk} - R_{1jk'}) \times \operatorname{sgn}(R_{2j'k} - R_{2j'k'}),$$

where $\operatorname{sgn}(a) = 1$ if $a > 0$ and 0 if $a < 0$. Finally, define Kendall's coefficient of concordance, W, among judges in Group 1, among judges in Group 2, and between judges in Groups 1 and 2, respectively, as

$$\hat{W}_g = \frac{12}{K(K^2 - 1)}\sum_{k=1}^{K}\left(\bar{R}_{g\cdot k} - \frac{K+1}{2}\right)^2$$

[1] Consistent with Hays (1960), in this article, we use Kendall's τ from Kendall and Smith (1939). Two other versions of Kendall's τ, τ_b and τ_c, which differ from τ in the handling of ties, are also commonly used.

$$= \frac{J-1}{J}\left[\frac{1}{\binom{J}{2}}\sum_{1\le j<j'\le J}\hat{r}_{s,g(jj')}\right]+\frac{1}{J}, \qquad g=1,2, \quad (1)$$

$$\hat{W}_{12} = \frac{12}{K(K^2-1)}\sum_{k=1}^{K}\left(\bar{R}_{1\cdot k}-\frac{K+1}{2}\right)\left(\bar{R}_{2\cdot k}-\frac{K+1}{2}\right)$$

$$= \frac{1}{J^2}\sum_{j=1}^{J}\sum_{j'=1}^{J}\hat{r}_{s,1(j)2(j')}. \quad (2)$$

Expression 1 is well-known; see, for example, Ehrenberg (1952). Expression 2 is derived in Appendix A. Therefore, Kendall's W is linearly related to the average of the pairwise Spearman's r_s.

Finally, define the population values of $\hat{r}_{s,1(jj')}$, $\hat{r}_{s,2(jj')}$, $\hat{r}_{s,1(j)2(j')}$ as $r_{s,1}$ $r_{s,2}$ $r_{s,12}$; the population values of $\hat{\tau}_{1(jj')}$, $\hat{\tau}_{2(jj')}$, $\hat{\tau}_{1(j)2(j')}$ as τ_1, τ_2, τ_{12}; and the population values of $\hat{W}_1$, $\hat{W}_2$, $\hat{W}_{12}$ as W_1, W_2, W_{12}.

Hays's (1960) method. For two groups with J judges in each group, Hays (1960) partitioned the average Kendall's τ between the rankings of all $\binom{2J}{2}$ pairs of judges into a within-group component and a between-groups component. The partition can be written as

$$\binom{J}{2}\bar{\hat{\tau}}_1+\binom{J}{2}\bar{\hat{\tau}}_2+J^2\bar{\hat{\tau}}_{12} \quad (3)$$

where $\bar{\hat{\tau}}_1$, $\bar{\hat{\tau}}_2$, $\bar{\hat{\tau}}_{12}$ represent, respectively, the average $\hat{\tau}$ between all pairs of judges in Group 1, between all pairs of judges in Group 2, and between all pairs of judges for which one was selected from Group 1 and another was selected from Group 2. Hays suggested analyzing the data using an analysis of variance (ANOVA), but no formal test was developed. It is conceivable that a statistic

$$T_{\text{Hays}} = \frac{\binom{J}{2}\bar{\hat{\tau}}_1+\binom{J}{2}\bar{\hat{\tau}}_2+J^2\bar{\hat{\tau}}_{12}}{\binom{J}{2}\bar{\hat{\tau}}_1+\binom{J}{2}\bar{\hat{\tau}}_2} \quad (4)$$

can be used. Using T_{Hays} is thus analogous to carrying out an ANOVA. The statistic T_{Hays} is sensitive to testing the hypotheses H_0: $\tau_1 \equiv \tau_2 = \tau_{12}$ vs. H_1: $\tau_{12} \neq \tau_1 \equiv \tau_2$. Under the null hypothesis, $T_{\text{Hays}} = 1$, whereas small values of T_{Hays} indicate a departure from concordance between groups.

Schucany and Frawley's (1973) and Li and Schucany's (1975) method. In a number of related works (Li & Schucany, 1975; Lyerly, 1952; Page, 1963; Schucany & Frawley, 1973), methods were considered that essentially used the average of the Spearman's r_s of the rankings between pairs of judges from the two groups. Both Lyerly (1952) and Page (1963) considered the situation where one of the two groups has $J > 1$ judges and the other group has $J = 1$ judge. In practice, this situation may arise in a study where the interest is to validate the rankings by the judges against some standard rankings. Lyerly (1952) suggested using the average Spearman's r_s of the rankings between the group with J judges and the group with a single judge. Page (1963) proposed summing the rankings across the J judges for each item and then calculating the correlation between the sums and the rankings by the sole judge in the other group.

Schucany and Frawley's (1973) statistic for measuring intergroup concordance for two groups, g and g', is

$$T_{\text{LSF1}} = \frac{\sum_{k=1}^{K}R_{g\cdot k}R_{g'\cdot k}-\frac{J^2K(K+1)^2}{4}}{\left[\frac{J^2(J-1)K^2(K+1)^2}{144}\right]^{1/2}}. \quad (5)$$

For large values of J and K, T_{LSF1} is approximately standard normal. Schucany and Frawley suggested using T_{LSF1} in the following way for testing intergroup concordance. Large positive values of T_{LSF1} would suggest concordance both within and between groups, values of T_{LSF1} near zero would indicate lack of concordance in either group, and large negative values of T_{LSF1} would suggest concordance within groups but disagreement between groups. Because T_{LSF1} is not restricted to the range of -1 to 1, Li and Schucany (1975) suggested transforming T_{LSF1} by $T'_{\text{LSF1}} = [J^2(K-1)]^{-1/2}T_{\text{LSF1}}$, which always lies within $[-1, 1]$. When the number of judges in one of the two groups is 1, T'_{LSF1} is identical to the statistic of Lyerly (1952) and T_{LSF1} is identical to the statistic of Page (1963; Li & Schucany, 1975, p. 419). Because these statistics are all linearly related, henceforth in this article, only T_{LSF1} will be studied further. It has been shown (Li & Schucany, 1975) that for two groups with J judges in each,

$$T'_{\text{LSF1}} = \bar{\hat{r}}_{s,12} = [J^2(K-1)]^{-1/2}T_{\text{LSF1}}, \quad (6)$$

where

$$\bar{\hat{r}}_{s,12} = \frac{\sum_{j=1}^{J}\sum_{j'=1}^{J}\hat{r}_{s,1(j)2(j')}}{J^2}$$

is the average Spearman's r_s between two judges: j from Group 1 and j' from Group 2. The hypotheses of Schucany and Frawley were H_0: $r_{s,1} \equiv r_{s,2} = r_{s,12} = 0$ vs. H_1: $r_{s,12} > 0$. However, Schucany (1978, p. 411) also suggested the possibility of using the statistic for testing the null hypothesis of complete concordance to an alternative of departure from concordance.

As pointed out by Li and Schucany (1975, p. 419), if the

two groups of judges are combined, then the average pairwise Spearman's r_s for the $\binom{2J}{2}$ pairs of judges in the combined group can be partitioned similarly to Partition 3, as

$$\binom{J}{2}\bar{\hat{r}}_{s,1} + \binom{J}{2}\bar{\hat{r}}_{s,2} + J^2 T'_{LSF1}. \tag{7}$$

Therefore, $\bar{\hat{r}}_{s,1}$, $\bar{\hat{r}}_{s,2}$ and T'_{LSF1} (or T_{LSF1}) can be interpreted as the within-group and between-group components, respectively, in an ANOVA of the Spearman's r_s in the combined sample. Based on Partition 7, an alternative statistic can be defined as follows:

$$T_{LSF2} = \frac{\binom{J}{2}\bar{\hat{r}}_{s,1} + \binom{J}{2}\bar{\hat{r}}_{s,2} + J^2 T'_{LSF1}}{\binom{J}{2}\bar{\hat{r}}_{s,1} + \binom{J}{2}\bar{\hat{r}}_{s,2}} = \frac{\binom{J}{2}\bar{\hat{r}}_{s,1} + \binom{J}{2}\bar{\hat{r}}_{s,2} + J^2\bar{\hat{r}}_{s,12}}{\binom{J}{2}\bar{\hat{r}}_{s,1} + \binom{J}{2}\bar{\hat{r}}_{s,2}}, \tag{8}$$

where the last equality is due to Expression 6. The statistic T_{LSF2} is therefore defined in the spirit of a test statistic of the interaction term in an ANOVA.

Hollander and Sethuraman's (1978) method. Hollander and Sethuraman (1978) advocated testing the null hypothesis of complete intergroup agreement versus the alternative of lack of complete intergroup agreement. Their method of analysis is based on the Mahalanobis distance between the average rankings of the first $K - 1$ items in the two groups of judges.[2] For group g, $g = 1, 2$; the average rankings of the judges in the group is $\bar{R}_g = (\bar{R}_{g\cdot 1}, \ldots, \bar{R}_{g\cdot K})$; and let

$$d_{kk'} = \frac{\sum_{g=1}^{2}\sum_{j=1}^{J}(R_{gjk} - \bar{R}_{\cdot k})(R_{gjk'} - \bar{R}_{\cdot k'})}{2J - 1}, \quad k, k' = 1, \ldots, K - 1.$$

Then Hollander and Sethuraman's statistic is

$$T_{HS} = \frac{J}{2}(\bar{R}_1 - \bar{R}_2)D^{-1}(\bar{R}_1 - \bar{R}_2)', \tag{9}$$

where D is the $(K - 1) \times (K - 1)$ matrix of $d_{kk'}$. The null hypothesis of agreement between groups can be rejected if T_{HS} is large. Hollander and Sethuraman suggested using the permutation distribution of T_{HS} for the purpose of calculating the significance level of the test. The statistic T_{HS} is designed for testing the hypotheses H_0: $r_{s,1} = r_{s,2} = r_{s,12}$ vs. H_1: $r_{s,12} < r_{s,1}, r_{s,2}$.

Kraemer's (1981) method. Kraemer (1981) suggested a conditional intergroup coefficient of concordance that, under the assumption of an equal number of judges between groups, is defined by

$$T_{Kra1} = \frac{\frac{12}{K(K^2-1)}\sum_{k=1}^{K}\left[\bar{R}_{\cdot k} - \frac{K+1}{2}\right]^2}{\frac{1}{2}\sum_{g=1}^{2}\frac{12}{K(K^2-1)}\sum_{k=1}^{K}\left[\bar{R}_{g\cdot k} - \frac{K+1}{2}\right]^2} = \frac{\frac{1}{4}(\hat{W}_1 + \hat{W}_2 + 2\hat{W}_{12})}{\frac{1}{2}(\hat{W}_1 + \hat{W}_2)}. \tag{10}$$

Using Expressions 1 and 2, T_{Kra1} can be written as

$$T_{Kra1} = \frac{\binom{J}{2}\bar{\hat{r}}_{s,1} + \binom{J}{2}\bar{\hat{r}}_{s,2} + J^2\bar{\hat{r}}_{s,12} + \frac{J}{2}}{\binom{J}{2}\bar{\hat{r}}_{s,1} + \binom{J}{2}\bar{\hat{r}}_{s,2} + \frac{J}{2}} \tag{11}$$

(see Appendix B for derivation). The expression on the right-hand side of Equation 11 is very similar to that in Equation 8 except for the quantity $J/2$ that appears in the numerator and denominator of the right-hand side of Equation 11. This quantity appears in Equation 11 because Kraemer's statistic includes all the pairwise correlations between different judges as well as the correlations in the same judge, which are all one. The right-hand side of Equation 11 is not a monotone transformation of the right-hand side of Equation 8 and, therefore, T_{LSF2} and T_{Kra1} behave quite differently, as will be seen in the simulation study that follows.

Kraemer's hypotheses are H_0: $[(W_1 + W_2 + 2W_{12})/4]/[(W_1 + W_2)/2] = 1$ vs. H_1: $[(W_1 + W_2 + 2W_{12})/4]/[(W_1 + W_2)/2] < 1$. If $r_{s,1} = r_{s,2}$, then Kraemer's hypotheses are equivalent to H_0: $W_{12} = W_1 = W_2$ vs. H_1: $W_{12} < W_1 = W_2$, which, because of Equation 2, are essentially the same as H_0: $r_{s,1} \equiv r_{s,2} = r_{s,12}$ vs. H_1: $r_{s,12} < r_{s,1} \equiv r_{s,2}$. Kraemer used a large sample theory approach and suggested a statistic based on the jackknife procedure (Arvesen, 1969; Quenouille, 1956):

[2] The method only uses the rankings of the first $K - 1$ items because the K rankings are related in the sense that the ranking of the Kth item is determined once the rankings of the first $K - 1$ items are known. Therefore, using the rankings in the first $K - 1$ items avoids singularity in the variance–covariance of the Mahalanobis distance statistic.

$$T_{\text{Kra2}} = \frac{1 - [(2J)T_{\text{Kra1}} - (2J-1)\bar{T}_{\text{Kra1}}]}{\frac{2J-1}{2J}\, s_{T_{\text{Kra1}}}} \qquad (12)$$

where

$$\bar{T}_{\text{Kra1}} = \sum_{g=1}^{2}\sum_{j=1}^{J} \frac{T_{\text{Kra1}}^{g(-j)}}{2J} \quad \text{and}$$

$$s_{T_{\text{Kra1}}} = \sqrt{\frac{\sum_{g=1}^{2}\sum_{j=1}^{J} (T_{\text{Kra1}}^{g(-j)} - \bar{T}_{\text{Kra1}})^2}{2J-1}}.$$

$T_{\text{Kra1}}{}^{g(-j)}$ is the value of T_{Kra1} calculated with the jth judge omitted from group g. Expression 12 can be compared with the t distribution with $J - 1$ degrees of freedom. In this article, alternative nonparametric tests based on T_{Kra1} and T_{Kra2} are constructed using permutation theory.

A method related to Kraemer's method is given by Legendre (2005), who considered the situation where one of the two groups has more than one judge and the other group has one judge. Legendre was primarily interested in post hoc tests to identify judges who are in disagreement with the others in a group. He considered calculating the average Spearman's r_s or the average Kendall's W between the group with one judge and all other judges in the other group and used permutation theory for hypothesis testing. Unlike the difference between T_{LSF2} and T_{Kra1}, the average Spearman's r_s and the average Kendall's W are equivalent (see Appendix A) and, therefore, the two statistics also give the same test results (see Legendre, 2005, p. 234).

Comparison of Methods

The following statistics—Hays (T_{Hays}), Li–Schucany–Frawley (T_{LSF1} and T_{LSF2}), Hollander–Sethuraman (T_{HS}), and Kraemer (T_{Kra1} and T_{Kra2})—were compared using a simulation study. The simulation was designed to mimic the situation where two groups of judges are each asked to rank a number of items in terms of their utility or preference. In the simulation study, the sampling distributions of the test statistics were determined using permutation theory. Permutation theory is desirable in the sense that no large sample approximation is required. Furthermore, when data are missing, most test statistics evaluated under large sample theory need to be modified accordingly, depending on how the missing entries are handled; permutation theory does not have this problem. Using simulations, Legendre and Lapointe (2004) and Legendre (2005) also demonstrated that permutation tests based on Kendall's W give greater statistical power than their counterparts based on large sample theory.

The rankings provided by the judges used in the simulations were induced using latent continuous random variables. This method of deriving rankings is consistent with Thurstone's (1927, 1931) theory of ranking choice alternatives, which states, among other things, that each alternative is based on a latent continuous utility value[3] that follows a normal distribution in the population of judges. Specifically, let the true utility value of item k among judges in group g be a_{gk}. In this formulation, the underlying utility of item k is identical for judges within the same group but it may be different from that for judges in a different group. When $a_{gk} = a_k$, then the underlying utility of each item is identical across groups. When $a_{gk} = a$, then all K items have the same underlying utility values, irrespective of groups. In this article, the latent utility value for item k in the jth judge in group g is written as

$$X_{gjk} = a_{gk} + e_{gjk}, \qquad (13)$$

where $e_{gjk} \sim N(0, \sigma_e)$ represents a measurement error in each judge that is independent of other judges, even for judges in the same group. With the latent utility values for the K items for the jth judge from group g defined as $X_{gj1}, \ldots, X_{gjK}$, the rankings for the K items, $\mathbf{R}_{gj} = (R_{gj1}, \ldots, R_{gjK})$ are the ranks associated with $(X_{gj1}, \ldots, X_{gjK})$ when they are ordered. In Model 13, (a_{1k}, a_{2k}) is modeled using a bivariate normal distribution with $M = (0, 0)$, $SD(a_{1k}) = SD(a_{2k}) = \sigma_a$, and $\text{corr}(a_{1k}, a_{2k}) = \rho$. In this case, the correlation (of the utility values) between two judges in the same group is

$$\rho_g = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_e^2}, \qquad g = 1, 2, \qquad (14)$$

and the correlation between two judges from different groups is

$$\rho_{gg'} = \frac{\rho\sigma_a^2}{\sigma_a^2 + \sigma_e^2}. \qquad (15)$$

Clearly, the relationship $0 \le \rho_{gg'} \le \rho_g \le 1$ holds in this case. Therefore, the induced concordance between judges from the same group is at least as high as that between judges from different groups. When $\rho = 1$, $a_{gk} = a_k$; therefore, the two groups are concordant.

A special situation is represented by $a_{gk} = a \sim N(0, \sigma_a)$. In this situation, the underlying utilities between the items are the same, irrespective of groups. In other words, on

[3] Kahneman (2000) identified two different definitions of utility: (a) experienced utility, which can be defined as a measure of the experience of pleasure or pain, and (b) decision utility, which is defined as the weight an individual places on an item among a list of alternatives. In this article, *utility* is simply defined as a measure of the relative value an individual places on an item among K items.

average, the judges have no opinions as to whether any one of the K items is better or worse than the others. This situation has been subjected to much debate in previous work (see discussions in Hollander & Sethuraman, 1978; Kraemer, 1981). However, it is unlikely that there would be interest in measuring concordance when judges have no opinions in the first place. Hence, we do not pursue this situation further.

In the simulation study, the different methods were compared under the hypotheses H_0: $c_1 \equiv c_2 = c_{12}$ vs. H_1: $c_{12} < c_1 \equiv c_2$, where c stands for a measure of concordance, which may be Kendall's τ or Spearman's r_s. The actual values of c_1, c_2, and c_{12} are induced by the structure of the data used in the simulations, and the simulations allow comparisons of methods using different measures of agreement (Kendall's τ or Spearman's r_s). The comparison of the methods then captures the sensitivities of the different methods to concordance between groups. This situation is similar to the familiar comparison of a Wilcoxon test and a *t* test of population location. The chosen set of hypotheses is a reasonable compromise between the suggestions of Hollander and Sethuraman (1978), Kraemer (1981), and Schucany (1978, p. 411) under the condition that the agreements within the two groups are identical ($c_1 \equiv c_2$). It is also similar to the formulation in Hays (1960), which assumed H_1: $c_{12} \neq c_1 \equiv c_2$. The use of H_1: $c_{12} < c_1 \equiv c_2$ here is reasonable because in practice it is difficult to find situations where the intergroup agreement is greater than the within-group agreement.

For each method studied, the critical value of the method's test statistic was calculated using permutation theory, as follows. Let $\mathbf{R} = (\mathbf{R}_{11}, \ldots, \mathbf{R}_{1J}; \mathbf{R}_{21}, \ldots, \mathbf{R}_{2J})$ be the rankings of the two groups of judges. Then under the null hypothesis that the two groups are concordant, all permutations of $\mathbf{R}$ are equally likely. Let T be one of the six statistics studied in this article: Then the $100 \times \alpha\%$ critical value for the permutation test is defined as T_α, where T_α is the lower $100 \times \alpha$ percentile[4] of the permutation distribution of T based on **R.** In the simulation study, T_α was approximated as follows. Vectors of **R** were simulated. For each **R** simulated, the values were randomly permuted and the statistic T was applied to the permuted vector. This process was repeated 10,000 times to simulate the permutation distribution of T, under each situation studied.

In the simulation study, two groups of judges were used. Each group had the same number of judges, J. Two different values of J were considered: 10 and 20. Two values of K were used: 5 and 10. Ranks were induced by the latent utility model (Equation 13) as described earlier. The values of the parameters were as follows: (a_{1k}, a_{2k}) follows a bivariate normal distribution with $M = (0, 0)$, $SD(a_{1k}) = SD(a_{2k}) = \sigma_a$, with $\sigma_a = 0.5, 0.75, 1$; $\text{corr}(a_{1k}, a_{2k}) = \rho$, with $\rho = 0, 1/3, 2/3, 1$; $e_{ijk} \sim N(0, 0.5)$. The choices of the values for σ_a and ρ were motivated by the following. The within-group correlation is given by Equation 14. Therefore, the value of σ_a represents the range of underlying utilities for the different items so that a larger value of σ_a results in larger differences in underlying utility values between items, leading to easier discrimination between items and, therefore, higher concordance within a group. The choices of $\sigma_a = 0.5, 0.75$, and 1 correspond to within-group correlations of .5, .7, and .8, respectively, which mimic moderate to strong within-group correlations. The degree of concordance between groups is governed by the parameter ρ. When $\rho = 0$, the rankings between groups are independent because $\rho_{12} = 0$. When $\rho = 1$, there is complete concordance between groups because in that case, $\rho_1 = \rho_2 = \rho_{12} = \sigma_a^2/(\sigma_a^2 + \sigma_e^2)$. The other cases ($\rho = 1/3$, 2/3) represent different degrees of departure from concordance.

For each combination of (J, K, σ_a, ρ), 10,000 simulation runs were used to compare the power of the different methods. A nominal Type I error rate of 5% was used for all methods.

Results

The results for $J = 10$ and $J = 20$ are similar and, therefore, for conservation of space, only results for $J = 10$ are presented. When $\rho = 1$, the null hypothesis is true and, therefore, the tests are expected to reject the null hypothesis at around the nominal Type I error rate (i.e., 5% here). The observed Type I errors of the six statistics—T_{Hays}, T_{LSF1}, T_{LSF2}, T_{HS}, T_{Kra1} and T_{Kra2} over 10,000 simulations—can be seen to be approximately the same as the nominal rate of 5% (see Table 1 and Figure 1, cases with $\rho = 1$). The top half of Figure 1 gives the results for $J = 10$ and $K = 5$, and the three plots give the power of tests using the six statistics for different values of σ_a and ρ. For each value of σ_a, the power of each test is plotted against decreasing values of ρ. If the underlying utility value of each item is a signal and the random variations in the judges' opinions are noises, then a larger value of σ_a represents a larger signal-to-noise ratio and, consequently, leads to greater statistical power to detect departure from concordance. This scenario is indeed the case, as can be seen: For each test, the power curve rises more rapidly toward 1 as σ_a increases from 0.5 to 1. For a fixed value of σ_a, the power curves of the tests using T_{LSF1}, T_{LSF2}, T_{HS}, T_{Kra1} and T_{Kra2} increase when the data move away from the null hypothesis of complete concordance (as represented by a decreasing value of ρ).

Unlike the tests using the other five statistics, the power curve of the test using T_{Hays} is not increasing monotonically as ρ decreases. The power is especially weak for situations

[4] For T_{HS}, $T = -T_{\text{HS}}$ because large values of T_{HS} indicate departure from H_0. The same convention was used for T_{Kra1}.

Table 1
Observed Type I Errors for Six Tests in Situations Represented in the Top Row of Figure 1

Method	σ_a		
	0.5	0.75	1
T_{Hays}	0.0462	0.0452	0.0485
T_{LSF1}	0.0512	0.0448	0.0473
T_{LSF2}	0.0478	0.0518	0.0454
T_{HS}	0.0523	0.0473	0.0474
T_{Kra1}	0.0504	0.0444	0.0473
T_{Kra2}	0.0498	0.0512	0.0485

Note. T_{Hays} = Hays statistic; T_{LSF1} and T_{LSF2} = Li–Schucany–Frawley statistics; T_{HS} = Hollander–Sethuraman statistic; T_{Kra1} and T_{Kra2} = Kraemer statistics.

with $\rho = 0$. This result is unexpected because, in theory, a smaller ρ should correspond to greater statistical power. However, the explanation may be found by studying the quantities $\bar{\hat{\tau}}_1$ and $\bar{\hat{\tau}}_2$ that appear in the denominator of T_{Hays}. These quantities are averages of Kendall's τ between pairs of judges, and Kendall's τ takes the value zero if the judges' rankings are independent. When $\rho = 0$, half of the judges have independent rankings from the other judges, and when the judges are permuted to find the permutation distribution of T_{Hays}, the independent rankings sometimes give zero (or near zero) values of $\bar{\hat{\tau}}_1$ and $\bar{\hat{\tau}}_2$, which in turn lead to very large values of T_{Hays} that affect the permutation distribution. This problem with Hays's method disappears with larger numbers of items because, in such cases, the chance of seeing the extreme situations described here becomes very small (see the bottom half of Figure 1). Finally, the loss of power has no effect on the Type I error of T_{Hays}, as demonstrated in Table 1, which shows the Type I error is approximately the same as the desired 5%.

Apart from the unusual phenomenon observed above, the best-performing tests are those using T_{LSF2}, T_{HS}, and T_{Kra2}, across all the scenarios studied. Overall, the best test statistic among the three is T_{LSF2}. For value of $K = 5$, T_{HS} is slightly better than T_{Kra2}, but for $K = 10$, the reverse is true. The test using T_{Kra1} performs much worse than that using T_{LSF2}, despite the similarities in the test statistics (see Expressions 8 and 11). The difference can be attributed to the fact that T_{Kra1} includes the nuisance correlations between rankings of the same judge, which add no value in drawing inferences because those correlations are always one. Using T_{Kra2} results in much better power than T_{Kra1} because the former can be seen as removing the nuisance effects of correlations within the same judge by pivoting (Beran, 1988). For $K = 5$, the performance of T_{Hays} is the poorest (see the top half of Figure 1). The test using T_{LSF1} also has low power across the range of values of ρ and σ_a. The reason for the unsatisfactory performance of T_{LSF1} can be explained by Expression 5, which shows that T_{LSF1} only uses the average between-groups Spearman's r_s and ignores the average within-group Spearman's r_s. As σ_a increases, the other three methods make use of the increased concordance within groups as benchmarks for detecting departure from concordance between groups, which leads to the increases in power of those three methods. However, T_{LSF1} does not use the within-groups information and hence the power curves are almost insensitive to the value of σ_a. (Compare the power curves of T_{LSF1} across the three plots in the top half of Figure 1.)

The results for $J = 10$, $K = 10$ (see the bottom half of Figure 1), are similar to those for $J = 10$, $K = 5$, except for the increase in power for all five methods across all scenarios. The increase in power for T_{Hays} is significant. This outcome is not surprising because the correlation between Kendall's τ and Spearman's r_s tends to unity when the number of items is large (Daniels, 1944; Kruskal, 1958).[5] The performance of the test using T_{Kra1} is also much improved compared with that for $K = 5$. The performance of T_{Kra1} is now almost as good as those of T_{Kra2}, T_{HS}, and T_{LSF2}. The improvement can be attributed to the fact that the relative influence of $J/2$ is less for a larger K, as $\bar{\hat{r}}_{s,1}$, $\bar{\hat{r}}_{s\,s,2}$, and $\bar{\hat{r}}_{s,12}$ are all increasing in K, whereas J is fixed for the same number of judges. Therefore, for fixed J, as K becomes large, T_{Kra1} becomes more similar to T_{LSF2} (see Expressions 8 and 11). Finally, the performance of the test using T_{LSF1} remains poor when compared with all other tests.

The results for $J = 20$ generally improve for all tests (with the exception of the test using T_{LSF1}), across the situations studied. For T_{Hays}, the problem with the loss of power almost completely disappeared. The results are available on request.

Missing Data

Missing data are a common problem in all empirical research (Little & Rubin, 2002). In a ranking experiment, missing data can result from a variety of reasons.[6] For example, when asked to rank a list of items from the most preferred to the least preferred, a judge may choose to rank only those items that he or she thinks are worth ranking, leaving all other items unranked. A judge may also leave out the rank of an item because he or she thinks the item is irrelevant, inapplicable, or unfamiliar. Finally, judges may

[5] The correlation between τ and r_s on two sets of K items is $2(K+1)/\sqrt{2K(2K+5)}$, which goes to 1 as K increases.

[6] One method to avoid missing data is to use a forced-choice paradigm, where judges are required to rank every single item. However, the ranks arising from a forced-choice paradigm may not always be meaningful. For example, if the judge is forced to give ranks to all K items, of which he or she is only familiar with a subset of items, then the rankings for those items that he or she is not familiar with may not be meaningful. Therefore, forced choice scales are not always desirable.

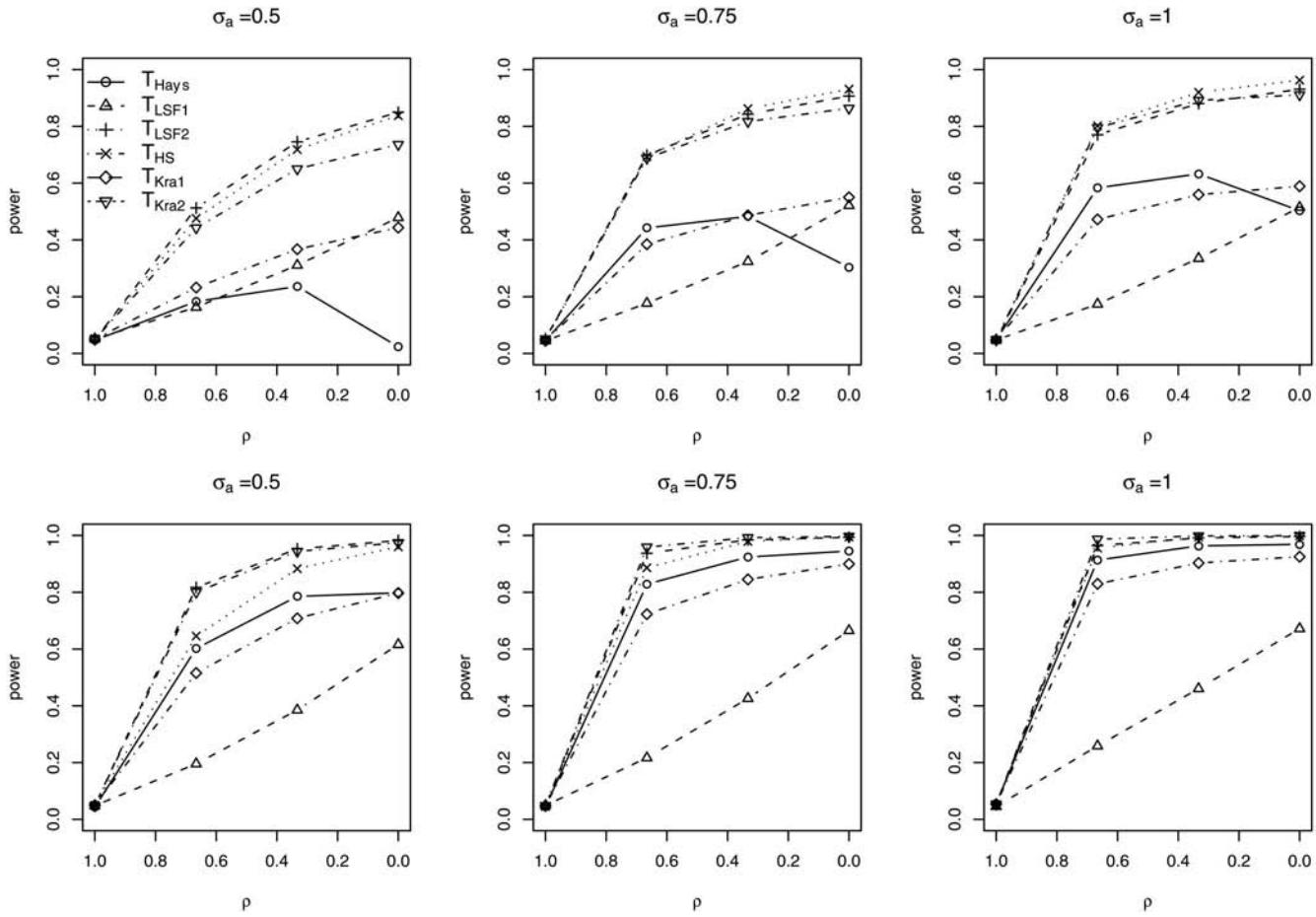

Figure 1. Power of five methods under various values of ρ and σ_a. For the three top grids, $J = 10$, $K = 5$. For the three bottom grids, $J = 10$, $K = 10$. T_{Hays} = Hays statistic; T_{LSF1} and T_{LSF2} = Li–Schucany–Frawley statistics; T_{HS} = Hollander–Sethuraman statistic; T_{Kra1} and T_{Kra2} = Kraemer statistics.

be asked to choose from a list of K items the top $K' < K$ items of their choice. In the first two cases, K' is random, but in the last case, K' is fixed. In any of these cases, it is reasonable to assume that the unranked items should receive a rank no higher than the last ranked item by that judge. Unranked items can also be a result of haphazardly missing data, in which case the missing ranks can be considered missing completely at random (Rubin, 1987). In ranking experiments, however, having items missing completely at random is less likely to be a possibility if the ranked items were given ranks without gaps.

In dealing with missing ranks in a ranking experiment, Schucany and Beckett (1976) and Stephens et al. (1977, 1978) proposed assigning all of the unranked items equal to one rank below the rank of the least preferred item. For example, if a judge only ranked a subset of the K items, giving rankings of 1 to K', $K' < K$, then the unranked items will all receive ranks of $K' + 1$. For a judge with missing rankings (that are missing completely at random), Alvo and Cabilio (1995) and Yu et al. (2002) considered all possible rankings that are compatible with the observed data. For example, if a judge gave rankings of (1, 2, 3, —) for four items, then the possible rankings that are compatible with the observed data are (1, 2, 3, 4), (1, 2, 4, 3), (1, 3, 4, 2), and (2, 3, 4, 1). (The last two configurations are possible because they do not change the relative rankings of the first three items as given by the judge.) Alvo and colleagues suggested calculating concordance measures with missing data by averaging the results using all possible observable (but not necessarily observed) rankings. However, even with moderate proportions of missing rankings, this method will involve an unmanageable number of possible rankings to consider. For example, with 10 items to be ranked by each judge, if 2 items are unranked by two judges, then the number of possible pairs of rankings in the two judges that are compatible with the observed data is (10!/8!)(2). If this process is repeated for all possible pairs of judges with missing rankings, the procedure will become infeasible to implement; for the same reason, it is next to impossible to study its properties.

A simulation study was conducted to study the behavior of the methods when unranked items are included in the data. Two missingness situations were considered. The first situation assumes that each judge only ranks those items that he or she considers important and that the unranked items simply take the tied rank one step below the last ranked item. This situation was created by assigning missing rankings to items with a latent utility value in the bottom 20% of the distribution of latent utility values. The second situation assumes that data are missing completely at random. For this situation, 20% of the latent utility values were randomly deleted and missing ranks are assigned to those items.

Two methods of imputing missing ranks were used in the simulation study. The first method assigns a rank equal to the tied rank one step below the last ranked item to all unranked items. The second method (Alvo & Cabilio, 1995; Yu et al., 2002) assumes that the unranked items are missing at random. As mentioned above, applying this method directly to a data set would require calculating the test statistic a large number of times, even when the percentage of unranked items is moderate. That is because under this method, the statistic is recalculated and then averaged over all possible rankings that are consistent with each incomplete set of rankings for each judge in each group.

Therefore, the following method for approximating the method of Alvo and Cabilio (1995) and Yu et al. (2002) was used. Suppose a judge gave a set of incomplete rankings of (1, 2, 3, —) for $K = 4$ items, then a random number U, from the continuous uniform $(0, K + 1 = 5)$ distribution is generated. U is compared with each of the items that have been ranked and a rank for the unranked item is imputed on the basis of the position of U among the other ranked items. Therefore, suppose U is 2.3: Then the imputed rank of the unranked item is 3 and the new rankings become (1, 2, 4, 3). Using this method, the imputed rankings will always be compatible with the original (incomplete) rankings, as suggested in Alvo and Cabilio (1995) and Yu et al. (2002). For each imputation method, the critical values for all the permutation tests were calculated using data imputed under that method. Therefore, all permutation tests automatically have the correct Type I error.

The missing data simulation study uses the same combinations of parameters (J, K, σ_a, and ρ) as in the complete data simulations. Once again, to conserve space, we focus the discussion on a selection of the results. Because T_{Hays}, T_{LSF1}, and T_{Kra1} are not competitive under the complete data simulations, they have been dropped from consideration in this part of the article.

Table 2 summarizes the power comparison between the complete data and missing data situations for $J = 10$, $K = 10$, when 20% of the data with the lowest underlying utility values are missing. The missing data were imputed using tied ranks one step below the last ranked item. The three columns under the heading Complete data in Table 2 give the powers of T_{LSF2}, T_{HS}, and T_{Kra2} in detecting H_1 when there are no missing data. The next three columns under the heading Missing data give their powers under the missing data situations. The last three columns give the ratio of the power with missing data to that with complete data. The results in Table 2 show that missing data lead to a substantial drop in performance in the methods across all scenarios tested. The drop in performance can be explained by the fact that when the missing data are imputed using tied ranks, the rankings between groups are more similar than they would be if the data were complete, leading to lower power. The power drop is greater for small values of σ_a and large values of ρ. In these situations, the utilities between items are similar and the concordance between groups is high. Therefore, the power is low even without missing data. When missing data are imputed by tied ranks, the power becomes even lower because the groups are even more difficult to discriminate. Among the three methods considered, T_{LSF2} is

Table 2
Comparison of Power Between Complete and Missing Data Situations

		Complete data[a]			Missing data[b]					
σ_a	ρ	T_{LSF2} (a)	T_{HS} (b)	T_{Kra2} (c)	T_{LSF2} (d)	T_{HS} (e)	T_{Kra2} (f)	d/a	e/b	f/c
0.50	0	0.982	0.960	0.975	0.483	0.392	0.351	0.492	0.408	0.360
	1/3	0.953	0.883	0.944	0.443	0.344	0.319	0.465	0.390	0.338
	2/3	0.817	0.646	0.800	0.300	0.230	0.197	0.368	0.357	0.246
0.75	0	0.994	0.995	0.997	0.836	0.764	0.749	0.841	0.768	0.752
	1/3	0.985	0.979	0.992	0.781	0.712	0.700	0.793	0.728	0.705
	2/3	0.937	0.886	0.959	0.627	0.555	0.535	0.669	0.626	0.558
1.00	0	0.997	0.999	0.999	0.944	0.926	0.921	0.947	0.927	0.922
	1/3	0.991	0.995	0.999	0.925	0.909	0.891	0.933	0.914	0.893
	2/3	0.964	0.956	0.987	0.837	0.806	0.791	0.867	0.843	0.802

Note. Missing data are created by deleting items with utility in the bottom 20% of the distribution. Missing data are imputed using tied ranks one step below the last ranked item. $J = 10$, $K = 10$.
T_{LSF2} = Li–Schucany–Frawley statistic; T_{HS} = Hollander–Sethuraman statistic; T_{Kra2} = Kraemer statistic.
[a]Power under complete data situation. [b]Power under missing data situation.

the most robust against missing data, followed by T_{HS}; the least robust is T_{Kra2}.

Table 3 gives results for $J = 10$, $K = 10$, when 20% of the data are randomly missing. The three columns under the heading Complete data are reproduced from Table 2 because the complete data and the two missing data situations were based on the same data. For the results in Table 3, the missing data were imputed by emulating the method of Alvo and Cabilio (1995) and Yu et al. (2002), as described earlier. Table 3 also shows a drop in performance in the methods, even though the drop in power is less severe than it is for comparable cases in Table 2. The more moderate power drop in this case can be attributed to the following. In the case of Table 2, the least important items are always given tied ranks. Because the least important items tend to be on one end of the distribution, the missing data are similar to censored observations, so items ranked low are almost always not observable. Because of that, the groups are concordant in those (unranked) items, leading to a greater chance of a nonsignificant test result. However, because the data in Table 3 are missing completely at random, all items are equally likely to be observed (or missing), and as long as the number of judges is not too small, the entire distribution of items is observable. The better results in Table 3 can also be attributed to the imputation method used. In this table, T_{LSF2} remains the best method, but the relative merits between T_{Kra2} and T_{HS} are reversed from those in Table 2.

Discussion

In this article, six statistics were considered for testing concordance in rankings between groups. Five of the statistics—Hays (T_{Hays}), the modified Li–Schucany–Frawley (T_{LSF2}), Hollander–Sethuraman (T_{HS}), and Kraemer (T_{Kra1}, T_{Kra2})—are analogous to carrying out an ANOVA on a particular measure of the overall agreement between rankings by the judges. T_{Hays} is a ratio of the between-group to within-group Kendall's τ, whereas T_{LSF2} is the parallel of T_{Hays} using Spearman's r_s. Kraemer's statistic, T_{Kra1}, is the ratio of the intergroup Kendall's W to the average intragroup Kendall's W, and T_{Kra2} is an adjusted version of T_{Kra1}. Finally, T_{HS} can be seen as a ratio of the between-group distance in the rankings of the K items to the within-group distance in the rankings of the same items. Because T_{LSF2}, T_{HS}, and T_{Kra2} all use ranks, it is not surprising that they perform similarly. Hays's statistic and T_{Kra1} are somewhat less satisfactory when the number of items is small. But in situations with a larger number of items, their performances approach those of T_{LSF2}, T_{HS}, and T_{Kra2}. Overall, the best-performing method is T_{LSF2}. However, T_{LSF1} performs poorly compared with the other statistics because it does not take the within-group concordance into consideration when making inferences.

We considered two different situations of missing data in this article. The performances of the methods are affected significantly when data are missing. The performance drop is more severe when the missing data arise from judges not ranking those items that they consider unimportant and the missing ranks are imputed using the average rank a step lower than the last ranked item. Among the three statistics considered, T_{LSF2}, T_{HS}, and T_{Kra1}, T_{LSF2} is the most robust. Other ways of imputing the missing ranks are possible. When a judge returned $K' < K$ ranks of K items, Critchlow (1985) suggested replacing the missing ranks with $(K + K' + 1)/2$. The idea is to make the mean imputed ranks the same as for a set of complete ranking of K items. Sen, Salama, and Quade (2003) advocated that the missing ranks should be imputed on the basis of some optimality criteria. Building on the suggestion of Sen et al. (2003), Salama and Quade (2004) considered weighting the items by their ranks, giving more emphasis to the top-ranked items. Their argu-

Table 3
Comparison of Power Between Complete and Missing Data Situations

		Complete data[a]			Missing data[b]					
σ_a	ρ	T_{LSF2} (a)	T_{HS} (b)	T_{Kra2} (c)	T_{LSF2} (d)	T_{HS} (e)	T_{Kra2} (f)	d/a	e/b	f/c
0.50	0	0.982	0.96	0.975	0.878	0.722	0.793	0.894	0.752	0.814
	1/3	0.953	0.883	0.944	0.749	0.571	0.665	0.786	0.646	0.705
	2/3	0.817	0.646	0.800	0.479	0.323	0.422	0.587	0.5	0.527
0.75	0	0.994	0.995	0.997	0.957	0.88	0.93	0.963	0.884	0.933
	1/3	0.985	0.979	0.992	0.886	0.769	0.868	0.9	0.785	0.875
	2/3	0.937	0.886	0.959	0.663	0.502	0.644	0.707	0.566	0.671
1.00	0	0.997	0.999	0.999	0.974	0.934	0.964	0.977	0.935	0.965
	1/3	0.991	0.995	0.999	0.927	0.844	0.923	0.935	0.848	0.924
	2/3	0.964	0.956	0.987	0.75	0.617	0.757	0.778	0.646	0.767

Note. Missing data are missing completely at random in 20% of the data. Missing data imputed using the method of Alvo & Cabilio (1995) and Yu et. al. (2002). $J = 10$, $K = 10$.
T_{LSF2} = Li–Schucany–Frawley statistic; T_{HS} = Hollander–Sethuraman statistic; T_{Kra2} = Kraemer statistic.
[a]Power under complete data situation. [b]Power under missing data situation.

ment is that the top-ranked items are the most important and, therefore, agreement in the top-ranked items should be given more weight. Similar work on weighting the Kendall's τ and Spearman's r_s appeared in Shieh (1998) and Shieh, Bai, and Tsai (2000), among others. In light of the substantial loss in power observed in the simulation study when items deemed unimportant are not ranked, it is worthwhile to consider these alternative methods of imputing the missing ranks. In certain cases, auxiliary information about the unranked items may be obtained and a method that effectively incorporates the auxiliary information may help to recover some of the lost information.

Another possible extension of the current study is to consider other missingness situations. One such situation is where the probability of a rank being missing is dependent on some observable covariates. Another situation that may be of interest is where the probability of a missing rank may depend on the value of the rank itself. These situations correspond to the missing at random and nonignorable missing data situations, respectively, as described by Rubin (1987). In both of these situations, correct specification of the missingness probability is crucial for valid inferences to be drawn.

Of further note, Thurstone's (1927, 1931) latent utility model is used to induce rankings in the simulations. Because the utility model is continuous by design, it will lead to rankings with no ties. In practice, there are invariably ties in some of the rankings. In the context of a latent utility model, ties in ranks result from items with small differences in underlying utilities, a situation that is reflected by a small value of σ_a in Model 13. As demonstrated in the simulations, a small value of σ_a leads to attenuation of power. Hays's method can also be expected to suffer under ties as ties are discarded from the calculation of Kendall's τ, on which Hays's method is based.

In the simulations, the three methods that use the actual rankings of the items (T_{LSF2}, T_{HS}, and T_{Kra2}) performed better than the method using pairwise comparison of ranks (T_{Hays}). When making inferences about the intergroup concordance, one must use both between- and within-group concordance. Failing to do so may lead to substantial loss in efficiency, as seen in the T_{LSF1}.

In the current study, we focused on situations with two groups of judges. The methods considered here can be generalized to situations with more than two groups of judges, if group membership is defined by a single factor. For the method of Kraemer, the numerator of Equation 10 is the (total) Kendall's *W* calculated on the basis of all possible pairs of judges from all groups, and the denominator is the average of the within-group Kendall's *W* (Kraemer, 1981, pp. 642–643). Similarly, for the methods of Hays (1960) and Schucany and Frawley (1973), the total Kendall's τ and total Spearman's r_s can be decomposed into within-group and between-group components, similar to an ANOVA (Beckett & Schucany, 1979), with no restriction on the number of groups. For the method of Hollander and Sethuraman (1978), a generalized version of the Mahalanobis distance can be defined, similar to a multivariate analysis of variance. It would be interesting to study the behavior of the methods for three or more groups of judges.

References

Alvo, M., & Cabilio, P. (1995). Rank correlation methods for missing data. *The Canadian Journal of Statistics, 23,* 345–358.

Arvesen, J. N. (1969). Jackknifing *U*-statistics. *Annals of Mathematical Statistics, 40,* 2076–2100.

Beckett, J., & Schucany, W. R. (1979). Concordance among categorized groups of judges. *Journal of Educational Statistics, 4,* 125–137.

Beran, R. (1988). Prepivoting test statistics: A bootstrap view of asymptotic refinements. *Journal of the American Statistical Association, 83,* 687–697.

Bonner, B. L. (2004). Expertise in group problem solving: Recognition, social combination, and performance. *Group Dynamics: Theory, Research, and Practice, 8,* 277–290.

Castel, A., Miró, J., & Rull, M. (2005). Validation of the faces pain scale in a sample of elderly Spanish individuals. *European Journal of Psychological Assessment, 21,* 265–270.

Critchlow, D. E. (1985). *Metric methods for analyzing partially ranked data.* Berlin, Germany: Springer-Verlag.

Daniels, H. E. (1944). The relation between measures of correlation in the universe of sample permutations. *Biometrika, 33,* 129–135.

Diaconis, P., & Graham, R. L. (1977). Spearman's footrule as a measure of disarray. *Journal of the Royal Statistical Society: Series B, 39,* 262–268.

Ehrenberg, A. S. C. (1952). On sampling from a population of rankers. *Biometrika, 39,* 82–87.

Elstein, A. S., Chapman, G. B., & Knight, S. J. (2005). Patients' values and clinical substituted judgments: The case of localized prostate cancer. *Health Psychology, 24*(4, Suppl.), S85–S92.

Feigin, P. D., & Alvo, M. (1986). Intergroup diversity and concordance for ranking data: An approach via metrics for permutations. *Annals of Statistics, 14,* 691–707.

Fisher, S. G., Macrosson, W. D. K., & Yusuff, M. R. (1996). Team performance and human values. *Psychological Reports, 79,* 1019–1024.

Hays, W. L. (1960). A note on average tau as a measure of concordance. *Journal of the American Statistical Association, 55,* 331–341.

Hollander, M., & Sethuraman, J. (1978). Testing for agreement between two groups of judges (with commentaries). *Biometrika, 65,* 403–411.

Kahneman, D. (2000). Experienced utility and objective happiness: A moment-based approach. In D. Kahneman & A. Tversky (Eds.), *Choices, values and frames* (pp. 673–692). Cambridge, United Kingdom: Cambridge University Press.

Kendall, M. G., & Gibbons, J. D. (1990). *Rank correlation methods* (5th ed.). London: Griffin.

Kendall, M. G., & Smith, B. B. (1939). The problem of *m* rankings. *Annals of Mathematical Statistics, 10,* 275–287.

Kraemer, H. C. (1981). Intergroup concordance: Definition and estimation. *Biometrika, 68,* 641–646.

Kruskal, W. H. (1958). Ordinal measures of association. *Journal of the American Statistical Association, 53,* 814–861.

Legendre, P. (2005). Species associations: The Kendall coefficient of concordance revisited. *Journal of Agricultural, Biological, and Environmental Statistics, 10*(2), 226–245.

Legendre, P., & Lapointe, F. J. (2004). Assessing congruence among distance matrices: Single-malt scotch whiskies revisited. *Australian & New Zealand Journal of Statistics, 46,* 615–629.

Li, L., & Schucany, W. R. (1975). Some properties of a test for concordance of two groups of judges. *Biometrika, 62,* 417–423.

Linhart, H. (1960). Approximate test for *m* rankings. *Biometrika, 47,* 476–480.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). Hoboken, NJ: Wiley.

Lohmann, A., Delius, J. D., Hollard, V. D., & Friesel, M. F. (1988). Discrimination of shape reflections and shape orientations by *Columba livia*. *Journal of Comparative Psychology, 102,* 3–13.

Lyerly, S. B. (1952). The average Spearman rank correlation coefficient. *Psychometrika, 17,* 421–428.

Mardia, K. V., Kent, J. T., & Bibby, J. M. (1979). *Multivariate analysis.* New York: Academic Press.

Marlowe, C. M., Schneider, S. L., & Nelson, C. E. (1996). Gender and attractiveness biases in hiring decisions: Are more experienced managers less biased? *Journal of Applied Psychology, 81,* 11–21.

McKnight, J., & Hills, A. M. (1999). Just how good are we at estimating attractiveness? *Psychology, Evolution and Gender, 1,* 213–238.

Miró, J., Huguet, A., & Nieto, R. (2005). Evaluation of reliability, validity, and preference for a pain intensity scale for use with the elderly. *Journal of Pain, 6,* 727–735.

Page, E. B. (1963). Ordered hypothesis for multiple treatments: A significant test for linear ranks. *Journal of the American Statistical Association, 58,* 216–230.

Pope, M., & Scott, J. (2003). Do clinicians understand why individuals stop taking lithium? *Journal of Affective Disorders, 74,* 287–291.

Quenouille, M. H. (1956). Notes on bias in estimation. *Biometrika, 43,* 353–360.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys.* New York: Wiley.

Rule, B. G., Bisanz, G. L., & Kohn, M. (1985). Anatomy of a persuasion schema: Targets, goals, and strategies. *Journal of Personality and Social Psychology, 48,* 1127–1140.

Salama, I. A., & Quade, D. (2004). Agreement among censored rankings using Spearman's footrule. *Communications in Statistics: Theory and Methods, 33,* 1837–1850.

Schucany, W. R. (1978). Comments on paper by M. Hollander and J. Sethuraman. *Biometrika, 65,* 410–411.

Schucany, W. R., & Beckett, J. (1976). Analysis of multiple sets of incomplete rankings. *Communications in Statistics: Theory and Methods, 5,* 1327–1334.

Schucany, W. R., & Frawley, W. H. (1973). A rank test for two group concordance. *Psychometrika, 38,* 249–258.

Sen, P. K., Salama, I. A., & Quade, D. (2003). Spearman's footrule under progressive censoring. *Journal of Nonparametric Statistics, 15,* 53–60.

Shieh, G. S. (1998). A weighted Kendall's tau statistic. *Statistics & Probability Letters, 39,* 17–24.

Shieh, G. S., Bai, Z., & Tsai, W. Y. (2000). Rank tests for independence—with a weighted contamination alternative. *Statistica Sinica, 10,* 577–593.

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology, 15,* 72–101.

Stephens, L. J., Claypool, P. L., & Buchalter, B. (1977). Partial ordering of populations. *Journal of Educational and Behavioral Statistics, 2,* 41–54.

Stephens, L. J., Claypool, P. L., & Buchalter, B. (1978). Partial ordering of populations [Correction notice]. *Journal of Educational and Behavioral Statistics, 3,* 384.

Stewart, A. E., & Stewart, E. A. (1996). A decision-making technique for choosing a psychology internship. *Professional Psychology: Research and Practice, 27,* 521–526.

Swanson, J. S., Wigal, S. B., & Udrea, D. (1998). Evaluation of individual subjects in the analog classroom setting: I. Examples of graphical and statistical procedures for within-subject ranking of responses to different delivery patterns of methylphenidate. *Psychopharmacology Bulletin, 34,* 825–832.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review, 79,* 281–299.

Thurstone, L. L. (1931). Rank order as a psychological method. *Journal of Experimental Psychology, 14,* 187–201.

Vidmar, G., & Cernigoj, M. (2004). Studying norms in small groups by means of multi-group concordance analysis. *Horizons of Psychology, 13*(4), 55–66.

Wanschura, R. G., & Dawson, W. E. (1974). Regression effect and individual power functions over sessions. *Journal of Experimental Psychology, 102,* 806–812.

Yu, P. L. H., Lam, K. F., & Alvo, M. (2002). Nonparametric rank tests for independence in opinion surveys. *Australian Journal of Statistics, 31,* 279–290.

(*Appendixes follow*)

Appendix A
Derivation of Expression 2

Write

$$\tilde{R}_{1jk} = R_{1jk} - \tfrac{1}{2}(K+1)$$

and

$$\tilde{R}_{2j'k} = R_{2j'k} - \tfrac{1}{2}(K+1).$$

$$\sum_{k=1}^{K}\sum_{j=1}^{J}\sum_{j'=1}^{J}(R_{1jk}-R_{2j'k})^2 = \sum_{k=1}^{K}\sum_{j=1}^{J}\sum_{j'=1}^{J}(\tilde{R}_{1jk}^2+\tilde{R}_{2j'k}^2$$

$$-2\tilde{R}_{1jk}\tilde{R}_{2j'k}) = J\sum_{k=1}^{K}\sum_{j=1}^{J}\tilde{R}_{1jk}^2 + J\sum_{k=1}^{K}\sum_{j'=1}^{J}\tilde{R}_{2j'k}^2$$

$$-\sum_{k=1}^{K}\sum_{j=1}^{J}\sum_{j'=1}^{J}2\tilde{R}_{1jk}\tilde{R}_{2j'k}$$

$$= J\sum_{j=1}^{J}\sum_{k=1}^{K}\left[R_{1jk}^2-(K+1)R_{1jk}+\frac{(K+1)^2}{4}\right]$$

$$+J\sum_{j'=1}^{J}\sum_{k=1}^{K}\left[R_{2j'k}^2-(K+1)R_{2j'k}+\frac{(K+1)^2}{4}\right]$$

$$-2\sum_{k=1}^{K}\sum_{j=1}^{J}\sum_{j'=1}^{J}\tilde{R}_{1jk}\tilde{R}_{2j'k}$$

$$= 2J^2\left[\frac{(2K+1)(K+1)K}{6} - \frac{K(K+1)^2}{4}\right]$$

$$-2\sum_{k=1}^{K}\sum_{j=1}^{J}\sum_{j'=1}^{J}\tilde{R}_{1jk}\tilde{R}_{2j'k} = J^2\frac{(K-1)(K+1)K}{6}$$

$$-2\sum_{k=1}^{K}\sum_{j=1}^{J}\sum_{j'=1}^{J}\tilde{R}_{1jk}\tilde{R}_{2j'k}. \quad \text{(A1)}$$

Using Equation A1 and the fact that $\bar{\hat{r}}_{s,12}$ is the average Spearman's r_s over all pairs of judges with one judge from each of Group 1 and Group 2,

$$\bar{\hat{r}}_{s,12} = \frac{1}{J^2}\sum_{j=1}^{J}\sum_{j'=1}^{J}\left[1-\frac{6}{K^3-K}\sum_{k=1}^{K}(R_{1jk}-R_{2j'k})^2\right]$$

$$= 1-\frac{6}{K^3-K}\frac{1}{J^2}\left[J^2\frac{(K-1)(K+1)K}{6}\right.$$

$$\left.-2\sum_{k=1}^{K}\sum_{j=1}^{J}\sum_{j'=1}^{J}\tilde{R}_{1jk}\tilde{R}_{2j'k}\right] = \frac{1}{J^2}\frac{12}{K^3-K}\sum_{k=1}^{K}\sum_{j=1}^{J}\sum_{j'=1}^{J}\tilde{R}_{1jk}\tilde{R}_{2j'k}$$

$$= \frac{1}{J^2}\frac{12}{K^3-K}\sum_{k=1}^{K}\left(\sum_{j=1}^{J}\tilde{R}_{1jk}\sum_{j'=1}^{J}\tilde{R}_{2j'k}\right)$$

$$= \frac{12}{K^3-K}\sum_{k=1}^{K}\left(\frac{1}{J}\sum_{j=1}^{J}R_{1jk}-\frac{K+1}{2}\right)\left(\frac{1}{J}\sum_{j'=1}^{J}R_{2j'k}-\frac{K+1}{2}\right)$$

$$= \frac{12}{K^3-K}\sum_{k=1}^{K}\left(\bar{R}_{1\cdot k}-\frac{K+1}{2}\right)\left(\bar{R}_{2\cdot k}-\frac{K+1}{2}\right) = \hat{W}_{12}.$$

Appendix B
Derivation of Expressions 10 and 11

The numerator of Expression 10 can be written as

$$\sum_{k=1}^{K}\left(\bar{R}_{\cdot k}-\frac{K+1}{2}\right)^2=\sum_{k=1}^{K}\left(\frac{\sum_{j=1}^{J}R_{1jk}+\sum_{j'=1}^{J}R_{2j'k}}{2J}-\frac{K+1}{2}\right)^2=\sum_{k=1}^{K}\left(\frac{1}{2J}\sum_{j=1}^{J}R_{1jk}-\frac{K+1}{4}+\frac{1}{2J}\sum_{j'=1}^{J}R_{2j'k}-\frac{K+1}{4}\right)^2=\sum_{k=1}^{K}\left[\left(\frac{1}{2J}\sum_{j=1}^{J}R_{1jk}-\frac{K+1}{4}\right)^2+\left(\frac{1}{2J}\sum_{j'=1}^{J}R_{2j'k}-\frac{K+1}{4}\right)^2+\left(\frac{1}{2J}\sum_{j=1}^{J}R_{1jk}-\frac{K+1}{4}\right)\times\left(\frac{1}{2J}\sum_{j'=1}^{J}R_{2j'k}-\frac{K+1}{4}\right)\right]=\frac{1}{4}\sum_{k=1}^{K}\left(\bar{R}_{1\cdot k}-\frac{K+1}{4}\right)^2+\frac{1}{4}\sum_{k=1}^{K}\left(\bar{R}_{2\cdot k}-\frac{K+1}{4}\right)^2+\frac{1}{2}\sum_{k=1}^{K}\left(\bar{R}_{1\cdot k}-\frac{K+1}{4}\right)\left(\bar{R}_{2\cdot k}-\frac{K+1}{4}\right)=\frac{1}{4}(\hat{W}_1+\hat{W}_2+2\hat{W}_{12}). \quad \text{(B1)}$$

The denominator of Expression 10 can be derived similarly.

Using Results 1 and 2, the last expression in Equation B1 can be written as

$$\frac{\frac{1}{4}\left(\frac{J-1}{J}\bar{\hat{r}}_{s,1}+\frac{1}{J}\right)+\frac{1}{4}\left(\frac{J-1}{J}\bar{\hat{r}}_{s,2}+\frac{1}{J}\right)+\frac{1}{2}\bar{\hat{r}}_{s,12}}{\frac{1}{4}\left(\frac{J-1}{J}\bar{\hat{r}}_{s,1}+\frac{1}{J}\right)+\frac{1}{4}\left(\frac{J-1}{J}\bar{\hat{r}}_{s,2}+\frac{1}{J}\right)}=\frac{\frac{J(J-1)}{2}\bar{\hat{r}}_{s,1}+\frac{J}{2}+\frac{J(J-1)}{2}\bar{\hat{r}}_{s,2}+\frac{J}{2}+J^2\bar{\hat{r}}_{s,12}}{\frac{J(J-1)}{2}\bar{\hat{r}}_{s,1}+\frac{J}{2}+\frac{J(J-1)}{2}\bar{\hat{r}}_{s,2}+\frac{J}{2}}=\frac{\binom{J}{2}\bar{\hat{r}}_{s,1}+\binom{J}{2}\bar{\hat{r}}_{s,2}+J^2\bar{\hat{r}}_{s,12}+J}{\binom{J}{2}\bar{\hat{r}}_{s,1}+\binom{J}{2}\bar{\hat{r}}_{s,2}+J}.$$