Journal of Wine Economics, Volume 10, Number 2, 2015, Pages 173–189
doi:10.1017/jwe.2015.18

Testing a Mixture of Rank Preference Models on Judges' Scores in Paris and Princeton*

Jeffrey C. Bodington [a]

Abstract

Rank preference and mixture models have been employed to evaluate the ranks assigned by consumers in taste tests of beans, cheese, crackers, salad dressings, soft drinks, sushi, animal feed, and wine. In many wine tastings, including the famous 1976 Judgment of Paris and the 2012 Judgment of Princeton, judges assign scores rather than ranks, and those scores often include ties. This article advances the application of ranking and mixture models to wine-tasting results by modifying the established use of a Plackett-Luce rank preference model to accommodate scores and ties. The modified model is tested and then employed to evaluate the Paris and Princeton wine-tasting results. Test results show that the mixture model is an accurate predictor of observed rank densities. Results for Paris and Princeton show that the group preference orders implied by the mixture model are highly correlated with the orders implied by widely employed rank-sum methods. However, the mixture model satisfies choice axioms that rank-sum methods do not, it yields an estimate of the proportion of scores that appear to be assigned randomly, and it also yields a preference order based on nonrandom preferences that tasters appear to hold in common. (JEL Classifications: A10, C00, C10, C12, D12)

Keywords: Mixture model, preference rank, statistics, wine tasting.

I. Introduction

Rank preference and mixture models have been applied to taste tests of snap beans (Plackett, 1975), cheese snacks (Vigneau et al., 1999), crackers (Critchlow, 1980), salad dressings (Theusen, 2007), soft drinks (Bockenholt, 1992), sushi (Chen, 2014), animal feed (Marden, 1995), an unidentified food (Cleaver and Wedel, 2001), and, now, recently, wine. Regarding wine, Bodington (2012) posited that observed wine-tasting results may have a mixture distribution with random, common

*The author thanks an anonymous reviewer for his or her helpful comments. All remaining errors and omissions are the responsibility of the author alone.
[a] Bodington & Company, 50 California Street #630, San Francisco, CA 94111; e-mail: jcb@bodingtonandcompany.com.

preference and idiosyncratic preference mixture components. Cao (2014) applied a mixture model with random-ranking and consensus-ranking components to the results of the 2009 California State Fair Commercial Wine Competition. Bodington (2015) applied a mixture of rank-preference models to the ranks assigned by experienced tasters during a blind tasting of Pinot Gris, and the results implied that common-preference agreement among tasters exceeded the random expectation of illusory agreement.

This article seeks to test and broaden the application of a rank-preference and mixture models to wine-tasting results expressed as numerical scores rather than ranks and to results that include ties between scores. The 1976 Judgment of Paris (Paris) and the 2012 Judgment of Princeton (Princeton, and together the Judgments) are well known and analyzed wine tastings that provide data and context for a replicable test of a mixture of rank-preference models that is modified to handle scores, with ties, that are converted to ranks. The judges' scores in Paris and Princeton had many ties.

The mixture of Plackett-Luce models applied to ranked data for a tasting of Pinot Gris in Bodington (2015) is summarized in Section II. Transforming numerical scores into ranks and choice axioms are discussed in Section III, and an addition to the mixture model to handle ties between numerical scores appears in Section IV. The resulting model satisfies the Luce and independence from irrelevant alternatives (IIA) choice axioms, it considers ties between scores, and it can be applied to the small sample sizes associated with most tastings. Next, in Section V, the model is tested on hypothetical data, on the Pinot Gris tasting results and with a Monte Carlo simulation. In Section VI, the mixture model is then applied to judges' scores for Paris and Princeton. The mixture model results yield estimates of potential Type I error, the proportion of tasters' scores that appear to be assigned randomly, and a preference order based on nonrandom preferences that tasters hold in common. That preference order is highly but not perfectly correlated with the implications of rank-sum methods applied to the Judgments, and, unlike rank-sum methods, it also complies with the Luce choice axiom and IIA. Conclusions follow in Section VII.

II. Mixture Model for Ranked Wines

A mixture distribution is the result of combining the distributions of two or more random variables. The distribution of the mixture is observable and the underlying component distributions may be unobservable or latent. A mixture model is a mathematical expression of the latent distributions and their observable combination. See McLachlan and Peel (2000), Mengersen et al. (2011) and References. As a starting point, notation and a mixture model for ranked wine-tasting results are summarized below.

The names of wines (each name w with a total of W wines) assessed by a taster (each taster t with a total of T tasters) are listed in an object vector $\mathbf{o} = (o_1, o_2, o_3, \ldots, o_W)$, and the respective scores assigned to the wines by each taster are listed in a score vector $\boldsymbol{x}_t = (x_{t,1}, x_{t,2}, x_{t,3}, \ldots, x_{t,W})$. When those scores are assigned a relative rank, or when a taster assigns ranks rather than scores, the result is a rank vector $\boldsymbol{r}_t = (r_{t,1}, r_{t,2}, r_{t,3}, \ldots, r_{t,W})$. Arranging the objects from most-preferred to least-preferred yields an order vector $\boldsymbol{y}_t = (y_{t,1}, y_{t,2}, y_{t,3}, \ldots, y_{t,W})$. Following Marden (1995) and others, ties between scores are indicated by their average rank and a superscript line over tied objects. Adapting notation from Kidwell et al. (2008), ■ symbolizes a one-unit interval between observed scores in a complete order vector $\boldsymbol{y}_t^c$. For example, assuming an object vector with four wines $\mathbf{o} = (A, B, C, D)$ and that a taster assigns scores $\boldsymbol{x}_t = (10, 15, 6, 10)$, the rank vector is $\boldsymbol{r}_t = (2.5, 1, 4, 2.5)$, the order vector is $\boldsymbol{y}_t = \left(B\overline{AD}C\right)$, and the complete order vector is $\boldsymbol{y}_t^c = \left(B\blacksquare\ \blacksquare\ \blacksquare\ \blacksquare\ \overline{AD}\blacksquare\ \blacksquare\ \blacksquare\ C\right)$.

Cao (2014) employed a mixture model with two latent classes of taster, those who appear to assign ranks randomly and those who appear to have consensus. Bodington (2015) analyzed two similar classes, employed a Plackett-Luce probability mass function (PMF, $f_t(\boldsymbol{y}_t|\boldsymbol{\rho})$) for the latent common-preference class of tasters, employed the mixture model in Equations (1) and (2), and tested that model on a blind tasting of Pinot Gris. The Plackett-Luce PMF appears in Equation (1) and ρ_i is the probability that wine i is selected as most preferred. See also Luce (1977) and Marden (1995). Next, a mixture model with two classes of taster expressing the probability of a taster's observed order vector ($f_t'(\boldsymbol{y}_t|\hat{\boldsymbol{\rho}}, \hat{\boldsymbol{\pi}})$) appears in Equation (2). In Equation (2), the probability of a taster's order vector $\boldsymbol{y}_t$ equals the probability that a taster assigns ranks randomly (π_r) times the random-ranking PMF, plus, the probability that a taster assigns ranks in accordance with common preferences (π_p) times the Plackett-Luce PMF. The π are known as mixture component weights or mixing proportions and a hat (^) indicates that the value of a parameter in the mixture model must be estimated.

$$f_t(\boldsymbol{y}_t|\boldsymbol{\rho}) = \prod_{i=1}^{W} \left(\frac{\rho_i}{\sum_{j=i}^{W} \left(\rho_j\right)} | \boldsymbol{y}_t, \boldsymbol{\rho} \right) \tag{1A}$$

$$0 \leq \rho_i \leq 1.0 \text{ and } 1.0 = \sum_{i=1}^{W} \rho_i \tag{1B}$$

$$f_t'(\boldsymbol{y}_t|\hat{\boldsymbol{\rho}}, \hat{\boldsymbol{\pi}}) = \hat{\pi}_r \cdot \left(\frac{1}{W!}\right) + \hat{\pi}_p \cdot f_t(\boldsymbol{y}_t|\hat{\boldsymbol{\rho}}) \tag{2A}$$

$$0 \leq \hat{\pi}_r \text{ and } \hat{\pi}_p \leq 1.0 \text{ and } 1.0 = \hat{\pi}_r + \hat{\pi}_p \tag{2B}$$

Several simple examples of mixture model results are worked by hand in Appendix A. As a check, those results also match the results of Equations (1) and (2). The

examples in Appendix A involve two wines and two to four tasters. Examples with more wines and tasters become intractable to solve by hand. For three wines and three tasters, the number of order vector combinations is $(3!)^3 = 216$. For ten wines and nine tasters, as there were in Paris and Princeton, the number of combinations is $(10!)^9 = 1.09 \times 10^{59}$.

Appendix A also provides examples of an important qualification regarding Bodington (2015) and this article. What are labeled here for convenience as assignments that appear to be random may actually themselves be a mixture of random assignments and idiosyncratic assignments that fit together to yield the characteristic flat shape of a uniform random distribution. As noted in the Conclusion, separating and identifying random and idiosyncratic assignments is work for the future. Neither of those are the common-preference assignments that are the focus of wine makers or the determinants of a tasting group's aggregate preference order.

Note again that the mixture model above applies to a tasting in which the protocol did not allow ties and tasters assigned ranks $\boldsymbol{r}_t$ rather than scores $\boldsymbol{x}_t$. Section III addresses transforming scores into ranks, and an addition in Section IV enables the model to handle ties between scores.

III. Transforming Scores into Ranks and Transitivity

Transforming numerical scores into ranks by sequentially ordering the scores is a customary practice. In application to the Judgments, Ashenfelter and Quandt (1999), Ginsburgh and Zang (2012), Quandt (2006, 2012), and Ward (2012) employed that transformation. Two aspects of that transformation and transitivity are discussed below.

First, although the transformation above does preserve transitivity, it may lose information. Continuing the example $\boldsymbol{x}_t = (10, 15, 6, 10)$ from Section II, the ranking transformation yields $\boldsymbol{y}_t = (\overline{BAD}C)$ but loses the extra information in $\boldsymbol{y}_t^c = (B\blacksquare\ \blacksquare\ \blacksquare\ \blacksquare\ \overline{AD}\blacksquare\ \blacksquare\ \blacksquare\ C)$. Not one judge in Paris or Princeton ranked her or his wines with ten consecutive scores. Intervals between scores of two and three were common, and most scored at least one wine next to an interval of four points. For example, in Paris, Pierre Brejoux's complete order vector was $\boldsymbol{y}_{PB}^c = (CB\blacksquare F\blacksquare IAD\blacksquare G\blacksquare HE\blacksquare\blacksquare\blacksquare\blacksquare J)$. Are the intervals between those scores random? Are they artifacts of flawed experimental design? Or, are the intervals information about the strengths of relative preferences?

Turning to the question of randomness, the skewness (γ_1) of tasters' scores in the Judgments shows that the intervals between scores do not appear to be random. Only two judges had symmetric distributions of scores with $\gamma_1 = 0$. Thirteen of the judges skewed right with $\gamma_1 > 0$, and the other 21 skewed left with $\gamma_1 < 0$. If that asymmetry in scores is caused by random interval ■ then the expectation of skewness is zero; $E(\gamma_1) = 0$. The mean skewness for the 9 * 4 = 36 observations in the

Judgments was −0.17, and the variance in that skewness was 0.27. The Students *t*-statistic for the hypothesis that $E(\gamma_1) = 0$ is $(0 - (-0.17))/\left(\sqrt{0.27}/\sqrt{36}\right) = 1.96$. That *t*-statistic has a one-tailed *p*-value of approximately 0.03. On that basis, the null hypothesis that the intervals ■ are random would be rejected for a large sample size but is marginal for a sample with 36 observations.

Next, flawed design can induce bias in any experiment. See Ashton (2014), Filipello (1955, 1956, 1957), Filipello and Berg (1958), Mantonakis et al. (2009), and Masson and Aurier (2015) for bias induced in wine-tasting results by the number of wines to be judged, the tasting protocol and tasters' expectations. See Cicchetti (2014) for a comparison of ranking and scoring the same wine. Another aspect of experimental design is the general guidance sometimes given to tasters about which characteristics in wine warrant certain score totals. For example, University of California at Davis and Jancis Robinson employ 20-point scales that assign levels of quality to point ranges. When judges score using a numerical point scale, do they assign scores independently according to each wine's quality? Or, is the scale merely a bound on judges' assessments of relative preference? Tasting experience implies that some judges assign a score to one wine according to a general zone of quality and then score the remaining wines "around" that anchor. Is the left skew found above an artifact of judges' expectations that the anchor score is a 15 to 17 out of 20 and then having more room for lower scores than for higher scores? This author is not aware of research that tests the reliability of score scales, how they may actually be used, and whether they induce any bias. See some of those issues discussed in Quandt (2012, p. 153) and Ward (2012, p. 159). On that basis, the notion that the experimental design of a tasting may induce ■ and bias in scores cannot be dismissed.

The possibility remains that differential intervals between numerical scores are information about the strengths of relative preferences. The probabilities in Plackett-Luce have been modified (from ρ_i to $\rho'_{t,i}$) by others to be functions of, or to make inferences about, additional information. In an application to betting on horse races, Benter (1994) added an exponent to capture information about the observed chance that a long-shot horse could win. In an application to election results, Gormley and Murphy (2007) made $\rho'_{t,i}$ an exponential function of each voter's notional distance from a candidate on various issues. In another analysis of election results, Gormley and Murphy (2008) made $\rho'_{t,i}$ a logistic function of the observed characteristics of voters. Applying those ideas to wine-tasting results seems perilous. Wine-tasting sample sizes are small, and there are few objectively measureable covariates. Moreover, disentangling strengths in relative preference from bias induced by experimental design is likely to be difficult with statistical significance, and it risks confusing the search for preferences that wine judges have in common.

In sum, transforming each judge's scores into ranks by merely ordering the scores does preserve the transitivity of preference for each judge. While that transformation may also lead to a loss of information, the value of that lost information appears to

be both speculative and intractable. Further analysis of that hypothesis may be more work for the future.

The second aspect of transitivity to be addressed here concerns comparing the ranks assigned by different judges to measure their aggregate, group, or social preference. Arrow's (1963) impossibility theorem concerning social choice and Luce's (1977) choice axiom both have implications based on transitivity about aggregating wine tasters' scores or ranks into measures of which wine or country "won," a preference order, and which wine or country "lost."

Arrow's theorem is considered here first. All of Ashenfelter and Quandt (1999), Ginsburgh and Zang (2012), Hulkower (2009), Quandt (2006,2012), and Ward (2012) compared the wines in the Judgments using various sums of judges' ranks (hereafter referred to as rank-sums methods). Quandt (2012, p. 153) explains that rank-sums methods violate a rule of choice logic sometimes known as independence from irrelevant alternatives (again, IIA). Quandt provides an example showing that excluding wine G from the white wines tasted in Princeton yields a different aggregate preference order for those that remain; $\mathbf{y} = (ADGBEIHFJC)$ changes to $\mathbf{y} = (\overline{AD}BEIHFJC)$. A simple example involves just two tasters and three wines. For $\boldsymbol{o} = (A, B)$, $\boldsymbol{r}_1 = (1, 2)$ and $\boldsymbol{r}_2 = (2, 1)$ the rank sums aggregate preference order is the tie $\mathbf{y} = (\overline{AB})$. Adding a wine C, for $\boldsymbol{r}_1 = (1, 2, 3)$ and $\boldsymbol{r}_2 = (3, 1, 2)$, the rank sums aggregate preference order is $\mathbf{y} = (CAB)$. Adding wine C thus changed the aggregate preference for A and B from indifference to preference. That example fails transitivity and demonstrates Arrow's general possibility theorem; there is no method of combining ranked individual expressions of preference into an aggregate that does not have logical inconsistencies. Restated without the double negative, every method of aggregating expressions of individual ranked preference into a measure of social preference has logical flaws (see also Marden [1995, p. 134]). IIA is one of the four criteria for Arrow's theorem, and it requires that aggregate social preference for an option A over option B should be independent of and should not be changed by individuals' preferences for A and B compared to option C.

Luce examined choice and IIA from a probabilistic perspective. See Luce (1977) for a formal statement and discussion of the Luce choice axiom (LCA). Applying the LCA to the simple example above, consider an urn containing equal numbers of balls marked A and B. Relative preference is the relative likelihood of drawing either A or B. With equal numbers of A and B balls, $\rho_A = \rho_B$ and the preference order is thus the same tie as above, $\mathbf{y} = (\overline{AB})$. Now, add just one ball marked C to the urn. In that case, $\rho_A = \rho_B > \rho_C$ thus $\mathbf{y} = (\overline{ABC})$. Next, add many more balls marked C such that $\rho_C > \rho_A = \rho_B$ and now $\mathbf{y} = (C\overline{AB})$. In both cases, transitivity is preserved, and IIA is obtained. In the formal statement of LCA, the aggregate relative preference for two objects depends on the ratio of their probabilities alone. In Luce (1977, p. 216), this is known as the constant ratio rule. The Plackett-Luce PMF

employed here in Equation (1) is consistent with the LCA (see Marden [1995, p. 134]; Plackett [1975]).

IV. Tied Scores

Ties are common in wine tastings. Every participant in Paris and Princeton assigned the same numerical score to at least two wines. Some of the judges assigned the same score to three or four wines. A taster may assign the same numerical score to two or more wines when he or she cannot distinguish between wines or decides that, all things considered, two or more wines deserve the same score. Quoting Quandt (2006, p. 9), "the option of using tied ranks enables tasters to avoid hard choices." An interpretation of ties from Kidwell et al. (2008) is that a taster needs more time or information than is available to actually differentiate between the tied wines.

Ties are often evaluated as the mean or expectation of the like objects. Spearman's rank correlation coefficient is calculated using the mean rank of tied objects. Kendall's rank correlation coefficient, the Mann-Whitney U test, and the Wilcoxon rank-sum test also employ the mean rank of tied objects. In a rank-preference model, it is convenient to treat the probability of an order vector containing ties as the expectation of the probabilities of the order vectors that are the permutations of the tie (each permutation m is $\boldsymbol{y}_{t,m}$ with a total of tp tie permutations). For example, the probability of $\boldsymbol{y}_t = (\overline{BADC})$ is the mean of the probabilities of $\boldsymbol{y}_{t,1} = (BADC)$ and $\boldsymbol{y}_{t,2} = (BDAC)$. See Critchlow (1980, p. 73), Kidwell et al. (2008, p. 1356), and Marden (1995, pp. 261, 269). The general form of that expectation appears in Equation (3).

$$E(f_t(\boldsymbol{y}_t|\hat{\boldsymbol{\rho}})) = \frac{1}{tp}\sum_{m=1}^{tp}\left(f_t(\boldsymbol{y}_{t,m}|\hat{\boldsymbol{\rho}})\right) \tag{3}$$

Although Equation (3) appears straightforward, it may not be easy in practice. In addition to the four wines that Odette Kahn scored in Paris as 12, she also assigned a score of 2 to two other wines. The number of her tie permutations tp is thus $4! \times 2! = 48$. In Princeton, Daniele Muelders assigned a score of 12 to four wines and a score of 15 to another four wines. The total of Muelders's tie permutations tp is $4! \times 4! = 576$.

Specific tests of the mixture model in Section V, including ties under Equation (3), show that the model is an accurate predictor of both observed rank densities and the mixture weights for random-behavior and common-preference score assignments.

V. Estimates, Tests, and Type I Error

The expectation maximization (EM) algorithm is a widely employed method of estimating the unknown parameters in mixture models. See Dempster et al. (1977),

McLachlan (2000), Mengersen et al. (2011), and References. In sum, EM iterates to climb a likelihood function. MATLAB code written by the author for the EM algorithm employed here, including an integrated maximum likelihood estimator (MLE), is available on request. Several tests of the mixture model in Equations (1) through (3), solved using the EM algorithm, are summarized below.

Test 1: A hypothetical 18 tasters rank six wines; six of those combine to have a random expectation, and 12 assign the same ranks to the same wines. The mixture weight for the random class should be 6/18 = 0.33. The EM solution does yield $\widehat{\pi}_r = 0.33$, and the estimates of $\widehat{\rho}_i$ imply the correct preference order.

Test 2: Again, 18 tasters rank six wines. Twelve of those combine to have a random expectation, and six assign the same ranks to the same wines. The mixture weight for the random class $\widehat{\pi}_r$ should be 12/18 = 0.67. The EM solution does yield $\widehat{\pi}_r = 0.67$, and the estimates of $\widehat{\rho}_i$ again imply the correct preference order. As an additional test, the model does replicate the ranked Pinot Gris results referenced above.

Test 3: The same data in Test 1 are evaluated now as scores rather than ranks. While the random class weight and density of the most preferred wine should be the same as in Test 1, the order should reverse because unity is the most-preferred rank, but it is the least-preferred score. The EM solution does yield those results. Further, the EM solution for changing three tasters' scores to prefer wine E the most and another three to prefer wine F the most does yield approximately $\widehat{\rho}_i = 0.5$ for the two wines that are tied for most-preferred.

Test 4: The data from Test 3 are evaluated again here except that, for the eighteenth taster alone, the scores on the most- and second-most-favored wines are tied. The total number of order vectors should be 17 + 2! = 19. Compared to the results for Test 3, the probability $\widehat{\rho}_i$ for the first-place wine should go down because that wine is now tied for second place for one taster. This is a test of Equation (3) for ties, and the EM solution does yield those results.

Test 5: In each Judgment, nine judges assigned scores between zero and 20 to each of ten wines. Accordingly, in this fifth test, each of nine tasters randomly assigns a score between zero and 20 to each of ten wines in a Monte Carlo simulation with 1,000 iterations. These draws include random ties and random intervals ■ between scores. For ten wines, the expected value of the probability that each wine is most-preferred $E(\widehat{\rho}_i)$ should be 1/10. That and other EM solution results are summarized in Table 1. Standard deviations (SD) are also reported. The expected results are obtained; none of the $E(\widehat{\rho}_i)$ are significantly different from 0.10, and their sum is close to unity. In addition, the expectation of the mixture weight for the class of tasters that appear to assign scores randomly $E(\widehat{\pi}_r)$ is 0.882. This implies that the probability of Type I error, false-positive illusory agreement on common preference, is approximately 1.000 − 0.882 = 0.118. See further discussion of Type I error and the tractable examples worked in Appendix A.

Table 1
Monte Carlo Simulation Results, 1,000 Iterations

	$\widehat{\rho}_i$		
	Mean	*SD*	*Other*
Wine			
A	0.103	0.279	
B	0.100	0.266	
C	0.105	0.297	
D	0.099	0.278	
E	0.097	0.267	
F	0.107	0.283	
G	0.095	0.274	
H	0.105	0.283	
I	0.100	0.278	
J	0.093	0.265	
Sum	1.003		
$\widehat{\pi}_r$			
Mean			0.882
SD			0.028
Log likelihood			
Mean			(125.19)
SD			1.2

VI. Application to Paris and Princeton

The title of this article portends a test of a mixture of rank-preference models using the scores that judges assigned in Paris and Princeton. Using the model presented and tested above, this section now presents that test.

A. Paris 1976

Eleven wine judges met at the Intercontinental Hotel in Paris, France, on May 24, 1976. Ten white wines and ten red wines were decanted into identical plain bottles. For each group of white and red wines, the tasting order was determined by drawing numbers from a hat. Each judge had two wine glasses, one for wine and the other for water. The protocol was step-by-step sequential; each judge tasted and scored a wine before the next wine was poured. Ten white wines were poured, then, following a break, ten reds. Each judged scored each wine on a scale of zero to 20, and the "official" ranking was based on the sums of the nine French judges' scores. See Taber (2005) and the author thanks both Mr. Taber and Mr. Spurrier for independently confirming this description of the protocol.

Among other sources, the judges' scores are available in Hulkower (2009, tables 3 and 8) and Lindley (2006). Spurrier and many others calculated overall preference using the total of scores for each wine. Ashenfelter and Quandt (1999) compared

the red wines using a rank-sums method. Quandt (2006) then compared the white wines using rank sums. Cicchetti (2006) evaluated both red and white wines using intraclass correlation coefficients to identify two classes of judges: those who ranked consistently with each other and those whose ranks appeared to be inconsistent. Hulkower (2009) reviewed the literature to date and presented another comparison using rank sums. Ginsburgh and Zang (2012) compared the red wines using rank sums and a game-theory–based measure of relative influence known as a Shapley value. See Olkin et al. (2015) for a recent survey of rank sums and other methods of analyzing wine-tasting data.

Mixture model results for Paris appear in Tables 2A and 2B. In addition to a preference order based on $\widehat{\rho}_i$, each table shows a preference order according to rank sums; most of the evaluations referenced above calculate rank sums and consider it superior to a sum of scores. For the white wines in Table 2A, the preference order implied by the mixture model is very close to that implied by rank sums. The correlation coefficient between the two is 0.94. The $\widehat{\rho}_i$ imply very strong agreement among judges on the first- and last-place wines and then much similarity in between. The estimate $\widehat{\pi}_r = 0.335$ implies that three of the nine judges appear to have assigned scores randomly. The likelihood ratio statistic (LRS), compared to the null hypothesis random-ranking Monte Carlo log likelihood result in Table 1, is $-2 * ((-125.19) - (-102.85) = 44.7$ and that LRS has a chi-square p- value < 0.001.

Table 2A
White Wines, 1976 Judgment of Paris

	Mixture model results			*Preference order*	
	$\widehat{\rho}_i$	$ln\ \widehat{\rho}_i$	*Other*	*Mixture model*	*Rank sums*
Château Montelena	0.997	(0.003)		1	1
Mersault Charmes	0.002	(6.177)		2	2
Chalone Vineyard	0.000	(8.404)		5	3
Spring Mountain	0.000	(8.398)		4	4
Freemark Abbey	0.000	(8.524)		6	6
Bâtard-Montrachet	0.000	(9.144)		7	7
Puligny-Montrachet	0.000	(9.837)		8	9
Beaune Clos des Mouches	0.000	(7.988)		3	5
Veedercrest	0.000	(10.080)		9	8
David Bruce	0.000	(48.209)		10	10
Sum	1.000				
$\widehat{\pi}_r$			0.335		
Log likelihood			(102.85)		
LRS (v. Monte Carlo)			44.7		

Judges' scores and rank-sums preference order from Borda results in Hulkower (2009).

Table 2B
Red Wines, 1976 Judgment of Paris

	Mixture model results			*Preference order*	
	$\widehat{\rho}_i$	$ln\ \widehat{\rho}_i$	*Other*	*Mixture model*	*Rank sums*
Stag's Leap	0.281	(1.269)		2	2
Château Mouton Rothschild	0.077	(2.565)		3	4
Château Haut Brion	0.043	(3.144)		4	1
Château Montrose	0.534	(0.627)		1	3
Ridge Monte Bello	0.002	(6.346)		7	5
Château Léoville	0.020	(3.925)		6	7
Mayacamas	0.041	(3.188)		5	6
Clos Du Val	0.000	(8.038)		10	8.5
Heitz Martha's Vineyard	0.001	(7.253)		8	8.5
Freemark Abbey	0.001	(7.585)		9	10
Sum	1.000				
$\widehat{\pi}_r$			0.451		
Log likelihood			(112.0)		
LRS (v. Monte Carlo)			26.4		

Judges' scores and rank-sums preference order from Borda results in Hulkower (2009).

Results for the red wines in Paris, in Table 2B, are similar to those for the white wines, however, the proportion of apparently random scoring as measured by $\widehat{\pi}_r$ increased. The preference orders implied by the mixture model and rank sums are again similar, and their correlation coefficient is 0.86. There is much general concordance, but specific agreement on only the second-place wine. The estimate of $\widehat{\pi}_r$ implies that four of the judges, one more than for the white wines, appear to have assigned scores randomly. That increase in the number of random-scoring judges could be due to palate fatigue, the characteristics of the red wines themselves, that announcement of results for the first flight and discussion among judges altered expectations, and other factors. The *p*-value of the LRS remains <0.001.

Finally, concerning Paris, the analysis above is intended only as a test of the mixture model on a famously available and analyzed set of data. The author does not intend to put forth any opinion about which wine or which country actually won or lost in Paris. Filipello (1955, 1956, 1957) and Filipello and Berg (1958) conducted various wine taste tests using sequential protocols and found evidence of primacy bias. Mantonakis et al. (2009) found evidence of both primacy and recency bias with an end-of-sequence protocol even among "high-knowledge" wine tasters. De Bruin (2005) examined singing and figure-skating competition results and found position bias in both step-by-step and end-of-sequence sequential

judging protocols. A step-by-step sequential protocol was employed in Paris, and the sequence of the wines has never been disclosed.[1]

B. Princeton 2012

Nine judges met on June 8, 2012, at Prospect House on the campus of Princeton University. Ten white wines and ten red wines were bagged and out of tasters' sight in another room. For each group of white and red wines, the tasting order was determined by drawing letters from a hat. In sharp contrast to the sequential protocol in Paris, in Princeton, there were ten glasses in front of each judge and each judge could taste and re-taste the wines in any order. Ten white wines were tasted and scored in a first flight, then ten red wines in a second. Water was available. Each judge scored each wine on a scale of zero to 20. See a description of the protocol and judges' scores in Ashenfelter and Storchmann (2012) and Taber (2012). Bodington (2012) showed that the "open" protocol employed in Princeton, in contrast to the sequential protocol employed in Paris, is not prone to sequential or flight position bias.

Mixture model results for Princeton and again the preference orders implied by rank sums appear in Tables 3A and 3B. Quandt (2012) compared the wines using rank sums. Ward (2012) calculated and offered caution about comparing the raw scores, ranks, centered, standardized, heterogeneous, and Friedman's test statistics. Ginsburgh and Zang (2012), as they did for Paris, compared the red wines using rank sums and Shapley values.

For the white wines in Princeton, results in Table 3A show that the preference orders implied by the mixture model and ranks sums are similar. The correlation coefficient between the two orders is 0.86. The estimates of $\widehat{\rho}_i$ for the common-preference class of taster imply their strong preference for first-place Clos des Mouches and strong preference against the last-place Ventimiglia. Those findings are consistent with those in Quandt (2012) and Ward (2012). In addition to preference-order implications that are similar to those of others, the mixture model implies substantial randomness in tasters' scores. The random mixture weight is $\widehat{\pi}_r = 0.670$, and the LRS is 0.2.

Results for the red wines in Princeton, shown in Table 3B, are similar to those for the white wines. There is general concordance between the mixture model and rank-sums results, and the correlation coefficient between the two orders is 0.78. The estimates of $\widehat{\rho}_i$ for the common-preference class of taster imply their strong preference

[1] Patricia Gallagher had a list with the tasting order and gave it to Taber (Taber, 2005, p. 200). In addition to a literature search, the author contacted George Taber, Steven Spurrier, an associate of Gallagher, and Jancis Robinson and posted an inquiry in the *Purple Pages* in an attempt to find the 1976 sequence. Other than that Puligny-Montrachet was the first white poured, both Spurrier (personal communication via e-mail, March 28, 2014) and Taber (personal communication via e-mail, April 22, 2014) conveyed to the author that the order is lost to history.

Table 3A
White Wines, 2012 Judgment of Princeton

	Mixture model results			*Preference order*	
	$\widehat{\rho}_i$	$ln\ \widehat{\rho}_i$	*Other*	*Mixture model*	*Rank sums*
Heritage	0.326	(1.122)		2	3
Unionville Pheasant	0.032	(3.444)		3	2
Puligny Montrachet	0.026	(3.660)		5	5
Clos des Mouches	0.567	(0.567)		1	1
Silver Decoy	0.031	(3.486)		4	4
Bellview	0.005	(5.278)		7	6.5
Ventimiglia	0.000	(19.044)		10	9
Mersault Charmes	0.006	(5.137)		6	10
Amalthea	0.004	(5.592)		9	8
Bâtard Montrachet	0.004	(5.507)		8	6.5
Sum	1.000				
$\widehat{\pi}_r$			0.670		
Log likelihood			(125.1)		
LRS (v. Monte Carlo)			0.2		

Rank-sums preference order from Quandt (2012).

Table 3B
Red Wines, 2012 Judgment of Princeton

	Mixture model results			*Preference order*	
	$\widehat{\rho}_i$	$ln\ \widehat{\rho}_i$	*Other*	*Mixture model*	*Rank sums*
Château Montrose	0.000	(15.834)		4	4
Château Mouton Rothschild	0.992	(0.008)		1	1
Silver Decoy	0.000	(17.617)		7	8
Heritage Estate	0.008	(4.817)		2	3
Bellview Lumière	0.000	(16.708)		5	7
Tomasello	0.000	(10.312)		3	5
Château Léoville	0.000	(18.323)		9	6
Amalthea Europa	0.000	(17.993)		8	9
Four JG's	0.000	(54.772)		10	10
Château Haut-Brion	0.000	(17.184)		6	2
Sum	1.000				
$\widehat{\pi}_r$			0.778		
Log likelihood			(123.0)		
LRS (v. Monte Carlo)			4.4		

Rank-sums preference order from Quandt (2012).

for Château Mouton Rothschild and strong preference against Four JG's. Again, that finding is consistent with those in Quandt (2012) and Ward (2012). As for the white wines, in addition to preference-order implications that are similar to those of others, the mixture model results also imply substantial randomness in tasters' scores. $\widehat{\pi}_r = 0.778$, and the LRS is 4.4. Again, as it did in Paris, $\widehat{\pi}_r$ increased for the second wine flight due to palate fatigue or other factors.

Standing back from the specific results and comparisons above, the test of the mixture model in Equations (1) through (3) using Paris and Princeton data has several implications. First, although the mixture model satisfies the Luce choice axiom that rank-sums methods do not, the mixture model's findings about aggregate preference order are similar but not identical to those of rank-sums methods. Second, in $\widehat{\pi}_r$, the mixture model implies and quantifies an increase in the appearance of random score assignments between the first and second wine flights that may be due to palate fatigue or other factors.

VII. Conclusion

A mixture of rank-preference models, including a Plackett-Luce PMF, was tested on the scores assigned by judges in the 1976 Judgment of Paris and the 2012 Judgment of Princeton. The aggregate- preference order implied by that mixture model complies with the Luce choice and IIA axioms, and it is also generally consistent with other published results. The mixture model has the added benefit of an estimate of Type I error, an estimate of the proportion of judge's scores that appear to be based on random-scoring behavior, and an estimate of judges' nonrandom common-preference order.

Although rank-preference and mixture models have been applied to taste tests of beans, cheese, crackers, salad dressings, soft drinks, sushi, and animal feed, application of the models to the unique challenges of wine-tasting results remains at an early stage. An analysis of alternatives to Plackett-Luce seems worthwhile. Restating the mixture to allow different proportions of common-preference agreement on different wines, changing π_p to $\pi_{p,i}$, seems realistic. While the mixture model presented above does parse observed tasting results into random and common-preference components, tasters' nonrandom but idiosyncratic preferences may account for much of the variance in observed scores. Some like more fruit than others, some like more acid, there are many examples. Future work is needed to identify and quantify those idiosyncratic preferences. Finally, the fundamental question of whether a zero to 20 point-score protocol induces ■ and bias in scores may be worth testing.

References

Arrow, K.J. (1963). *Social Choice and Individual Values*. 2nd ed. New York: John Wiley & Sons.

Ashenfelter, O., and Quandt, R.E (1999). Analyzing a wine tasting statistically. *Chance*, 12, 16–20.

Ashenfelter, O., and Storchmann, K. (2012). Editorial: The Judgment of Princeton and other articles. *Journal of Wine Economics*, 7(2), 139–142.
Ashton, R.H. (2014). Nothing good ever came from New Jersey: Expectations and the sensory perception of wine. *Journal of Wine Economics*, 9(3), 304–319.
Benter, W. (1994). Computer-based horse race handicapping and wagering systems: A report. In W.T. Ziemba, V.S. Lo, and D.B. Haush (eds.), *Efficiency of Racetrack Betting Markets*. San Diego: Academic Press, 183–198.
Bockenholt, U. (1992). Thurstonian representation for partial ranking data. *British Journal of Mathematical and Statistical Psychology*, 45, 31–49.
Bodington, J. (2012). 804 Tastes: Evidence on randomness, preferences and value from blind tastings. *Journal of Wine Economics*, 7(2), 181–191.
Bodington, J. (2015). Evaluating wine-tasting results and randomness with a mixture of rank preference models. *Journal of Wine Economics*, 10(1), 31–46.
Cao, J. (2014). Quantifying randomness versus consensus in wine quality ratings. *Journal of Wine Economics*, 9(2), 202–213.
Chen, W. (2014). *How to Order Sushi*. PhD dissertation, Harvard University.
Cicchetti, D.V. (2006). The Paris 1976 wine tasting revisited once more: Comparing ratings of consistent and inconsistent tasters. *Journal of Wine Economics*, 1(2), 125–140.
Cicchetti, D.V. (2014). Blind tasting of South African wines: A tale of two methodologies. American Association of Wine Economists. AAWE Working Paper No. 164.
Cleaver, G., and Wedel, M. (2001). Identifying random-scoring respondent in sensory research using finite mixture regression results. *Food Quality and Preference*, 12, 373–384.
Critchlow, D.E. (1980). *Metric Methods for Analyzing Partially Ranked Data*. New York: Springer.
De Bruin, W. (2005). Save the last dance for me: Unwanted serial position effects in jury evaluations. *Acta Psychologica*, 118(3), 245–260.
Dempster, A.P., N.M. Laird, and D.B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B*, 39(1), 1–38.
Filipello, F. (1955). Small panel taste testing of wine. *American Journal of Enology*, 6(4), 26–32.
Filipello, F. (1956). Factors in the analysis of mass panel wine-preference data. *Food Technology*, 10, 321–326.
Filipello, F. (1957). Organoleptic wine-quality evaluation II: Performance of judges. *Food Technology*, 11, 51–53.
Filipello, F., and H.W. Berg (1958). The present status of consumer tests on wine. Paper presented at the Ninth Annual Meeting of the American Society of Enologists, Asilomar, Pacific Grove, California, June 27–28.
Ginsburgh, V., and Zang, I. (2012). Shapley ranking of wines. *Journal of Wine Economics*, 7 (2), 169–180.
Gormley, I.C., and Murphy, T.B. (2007). A latent space model for rank data. In E. Airoldi, D. M. Blei, S.E. Fienberg, A. Goldenberg, E.P. Xing, and A.X. Zheng (eds.), *Statistical Network Analysis: Models, Issues and New Directions*. Berlin: Springer, 90–102.
Gormley, I.C., and Murphy, T.B. (2008). A mixture of experts model for rank data with applications in election studies. *Annals of Applied Statistics*, 2(4), 1452–1477.
Hulkower, N.D. (2009). The Judgment of Paris according to Borda. *Journal of Wine Research*, 20(3), 171–182.
Kidwell, P., Lebanon, G., and Cleveland, W.S. (2008). Visualizing incomplete and partially ranked data. *IEEE Transactions on Visualization and Computer Graphics*, 14(6) 1356–1364.
Lindley, D.V. (2006). Analysis of a wine tasting. *Journal of Wine Economics*, 1(1), 33–41.

Luce, R. D. (1977). The choice axiom after twenty years. *Journal of Mathematical Psychology*, 15 (3), 215–233.

Mantonakis, A., Rodero, P., Lesschaeve, I., and Hastie, R. (2009). Order in choice: Effects of serial position on preferences. *Psychological Science*, 20(11), 1309–1312.

Marden, J.I. (1995). *Analyzing and Modeling Rank Data*. London: Chapman & Hall.

Masson, J., and Aurier, P. (2015). Should it be told or tasted? Impact of sensory versus non-sensory cues on the categorization of low-alcohol wines. *Journal of Wine Economics*, 10(1), 62–74.

McLachlan, G., and Peel, D. (2000). *Finite Mixture Models*. New York: John Wiley & Sons.

Mengersen, K.L., Robert, C.P., and Titterington, D.M. (2011). *Mixtures: Estimation and Application*. New York: John Wiley & Sons.

Olkin, I., Lou, Y., Stokes, L. and Cao, J. (2015). Analyses of wine-tasting data: A tutorial. *Journal of Wine Economics*, 10(1), 4–30.

Plackett, R.L. (1975). The analysis of permutations. *Applied Statistics*, 24(2), 193–202.

Quandt, R.E. (2006). Measurement and inference in wine tasting. *Journal of Wine Economics*, 1(1), 7–30.

Quandt, R.E. (2012). Comments on the Judgment of Princeton. *Journal of Wine Economics*, 2 (7), 152–154.

Taber, G.M. (2005). *Judgment of Paris: California vs. France and the Historic 1976 Paris Tasting That Revolutionized Wine*. New York: Scribner.

Taber, G.M. (2012). The Judgment of Princeton. *Journal of Wine Economics*, 2(7), 143–151.

Theusen, K.F. (2007). Analysis of ranked preference data. Master's thesis, Technical University of Denmark, Kongens Lyngby, Denmark.

Vigneau, E., Courcoux, P., and Semenou, M. (1999). Analysis of ranked preference data using latent class models. *Food Quality and Preference*, 10(3), 201–207.

Ward, D.L. (2012). A graphical and statistical analysis of the Judgment of Princeton wine tasting. *Journal of Wine Economics*, 7(2), 155–168.

Appendix A: Examples of Mixture Model Results and Type I Error

Several simple examples of mixture model results and Type I error appear below. Examples with two wines and two to four tasters are tractable to work out by hand as shown. As a check, the EM solutions to Equations (1) and (2) do match the results below.

Wines/ Tasters	*Combinations of permutations*	*Observed probability A is first*	$\widehat{\rho_A}$	$\widehat{\pi_r}$	$E(\widehat{\pi_r})$	*Type I Error E* $(\widehat{\pi_p})$
	AB, AB	1.00	1.00	0.00		
2/2	AB, BA	0.50	0.50	1.00	2/4 = 0.50	0.50
	BA, AB	0.50	0.50	1.00		
	BA, BA	0.00	0.00	0.00		
	AB, AB, AB	1.00	1.00	0.00		
2/3	AB, AB, BA	0.67	1.00	0.67	4/8 = 0.50	0.50
	AB, BA, AB	0.67	1.00	0.67		
	AB, BA, BA	0.33	0.00	0.67		
	BA, AB, AB	0.67	1.00	0.67		
	BA, AB, BA	0.33	0.00	0.67		
	BA, BA, AB	0.33	0.00	0.67		
	BA, BA, BA	0.00	0.00	0.00		
	AB, AB, AB, AB	1.00	1.00	0.00		
2/4	AB, AB, AB, BA	0.75	1.00	0.50	10/16 = 0.625	0.375
	AB, AB, BA, AB	0.75	1.00	0.50		
	AB, AB, BA, BA	0.50	0.50	1.00		
	AB, BA, AB, AB	0.75	1.00	0.50		
	AB, BA, AB, BA	0.50	0.50	1.00		
	AB, BA, BA, AB	0.50	0.50	1.00		
	AB, BA, BA, BA	0.25	0.00	0.50		
	BA, AB, AB, AB	0.75	1.00	0.50		
	BA, AB, AB, BA	0.50	0.50	1.00		
	BA, AB, BA, AB	0.50	0.50	1.00		
	BA, AB, BA, BA	0.25	0.00	0.50		
	BA, BA, AB, AB	0.50	0.50	1.00		
	BA, BA, AB, BA	0.25	0.00	0.50		
	BA, BA, BA, AB	0.25	0.00	0.50		
	BA, BA, BA, BA	0.00	0.00	0.00		

Note that if the observed distribution has the exact "flat" form of a random distribution, the estimates of $\widehat{\pi_r}$ and $\widehat{\rho_i}$ are indeterminate because the observed distribution of ranks in that case is homogeneous and thus does not appear to be a mixture. The author has never observed that pattern in actual tasting results. See also discussion of Type I error in Bodington (2015).