Journal of Wine Economics (2025), 1–14
doi:10.1017/jwe.2025.10093

CAMBRIDGE UNIVERSITY PRESS

ARTICLE

Wine Stars & Bars: The combinatorics of critic consensus and the usefulness of order preference models

Jeff Bodington

Bodington & Co, San Francisco, CA, USA
Email: jcb@bodingtonandcompany.com

Abstract

Sums of the ratings that judges assign to wines are a near universal method of determining the winners and losers of wine competitions. Sums are easy to calculate and easy to communicate, but seven flaws make sums of ratings a perilous guide to relative quality or preference. Stars & Bars combinatorics show that the same sum can be the result of billions of compositions of ratings and that those compositions, for the same sum, can contain dispersion that ranges from universal consensus to apparent randomness to polar disagreement. Order preference models can address both order and dispersion, and an example using a Plackett–Luce model yields maximum likelihood estimates of top-choice probabilities that are a defensible guide to relative quality or preference.

Keywords: wine; ratings; statistics; combinatorics; consensus

JEL classifications: C; C4; C46

I. Introduction

International Organization of Wine and Vine (OIV) rules prescribe that the relative merits of wines entered in a competition are determined by the sums of the ratings that judges assign to the wines. That same procedure was employed in the 1976 Judgment of Paris, and it continues to be employed in hundreds of wine competitions around the world. A difficulty with sums of ratings is that a deep literature shows that wine ratings are stochastic, sample sizes are small, the range effect biases judge influence, converting scores to ranks ignores information, some judges are more reliable than others, judges may not apply uniform standards, and a sum is not an aggregate utility. Those findings make the sums or ratings unreliable guides to relative merit and consensus among judges.

While wine ratings are not merely random, they are stochastic and literature concerning that nature of ratings and sums of ratings is summarized in Section II. Application of a combinatoric binomial formula known as Stars & Bars in Section III

shows that many billions of combinations of ratings have the same sum and that uncertain ratings can have many different sums. Section IV shows that dispersion among the ratings that compose a sum must be considered but is also difficult to consider. That finding supports, in Section V, application of a Plackett–Luce order preference model that considers dispersion and that yields maximum likelihood estimates of top-choice probabilities that are a defensible guide to relative quality or preference.

Ratings assigned by judges to white wines at the 1976 Judgement of Paris reported in De Nicoló (2025, p. 23) are employed as an example. That data and MATLAB code for the figures and tables in this article are available on request.

II. Sum conundrums

OIV (2021, p. 6) prescribes that at least five judges shall each assign a score to a wine that is the sum of characteristic-specific sub-scores (p. 13), and the overall rating for each wine is the average of the judges' scores (p. 14). An average is of course merely a ratio of a sum. The overall ratings at the Judgment of Paris were sums of nine judges' scores, the overall ratings assigned in the California State Fair Commercial Wine Competition (CSF) are sums of three to five judges' scores, Kopsacheilis *et al.* (Kopsacheilis et al., 2024, p. 291) reported on averages of crowd-sourced ratings, and there are hundreds of other national, state or province, county or prefecture, and publication wine assessments that employ sums of ratings to express the absolute or relative merits of wines.

Sums of observed ratings are easy to calculate and easy to communicate, but they leave several issues unresolved as follows:

(1) Ratings are stochastic. Although rating assignments are not merely random, the deep literature surveyed in J. Bodington (2022) shows that ratings are uncertain and affected by physiochemical, sensory, and cognitive biases. A rating observed is one draw from a latent distribution that is wine- and judge-specific. If ratings are stochastic then sums of ratings are also stochastic.

(2) Sample sizes are small. The Paris and CSF examples above involved three to nine judges and blind replicates within flights are extremely rare so the sample size per-judge-per-wine is usually one. The Law of Large Numbers (LLN) states that the sample mean of independent and identically distributed (IID) random variables tends toward the expectation of those IID variables as the size of the sample tends toward infinity. Judges' ratings may be independent (I) but they are heteroscedastic and not identically distributed (not ID). Thus, a small-sample sum of judges' ratings may or may not be close to the actual expectation of the sum.

(3) The range effect biases influence. Ashenfelter and Quandt (1999) and others have pointed out that a judge who assigns scores within a broad range has more effect on relative sums than a judge who assigns ratings within a narrow range. While differences in range may reflect differences in the intensity of judges' opinions that range effect also biases the relative qualities or preferences that sums imply about wines.

(4) Ranks ignore differential intensity. To address the range effect in #3 above, Ashenfelter and Quandt (1999) and others suggest, when scores are assigned, transforming those scores into ranks for the purpose of calculating relative preference. Tinsely & Weiss (1975) counsel against converting scores to ranks because only order is preserved and information about differential intensity of preference is lost. J.C. Bodington (2015) examined the skewness in scores assigned to wines and concluded that differential intervals between scores do not appear to be random. Considering that result, transforming scores to ranks does appear to ignore information about the relative intensity of judges' rating assignments.

(5) Some judges are more consistent than others. J. Bodington (2022), cited above, summarized results published by those who have employed blind replicates to show that some judges are more reliable than others. He showed that the cross-correlations between the scores that judges assign to the same wines are on average positive but that about 10% of judges assign scores that are indistinguishable from random assignments. Kahneman *et al.* (Kahneman et al., 2021, p. 80–86, 215–258) describe variance in wine ratings and other areas of human judgment including physicians' diagnoses, radiologists' assessments of x-rays, forensic experts' fingerprint identifications, and judges' sentencings of criminals. Those findings imply a corollary to the famous test advanced by mathematician Alan Turing. Turing (1950) proposed that a computer can be described as an intelligent machine if, in a typewritten conversation, a computer can imitate a human so well that the computer and human responses are indistinguishable. A corollary to that test suggested here is that if the ratings a wine judge assigns are indistinguishable from those assigned by a random number generator, then that judge can be described as a random number generator. There is no justification for giving a random number generator any influence on wine competition results.

A simple sum of ratings gives equal weight to every judge regardless of differences in judges' consistencies. Cochran (1937) showed that a simple sum of independently distributed random variables with different variances is not a minimum variance estimate of the sum. He derived what is now known as the inverse variance rule to calculate a mean or sum that is a minimum variance estimate. However, in spite of differences in judges' consistencies, judges do sometimes reach unanimous conclusions. That implies that differences in consistency are not constant. At a hypothetical convergence, all judges could agree that a blind taste of 100% vinegar would rate lower than all other wines. Similarly, all judges could rate one wine in a flight highest if all the other wines are 100% vinegar. That notion is posed here as the vinegar axiom; even low-skill judges can identify a very flawed wine and converge in agreement with high-skill judges.

(6) The definition of and adherence to rating standards is opaque. See for example Circle of Wine Writers (2025). In discussion among judges after a tasting, judges find that one judge prioritized typicity, another prioritized quality, and yet another prioritized "I liked it." Several judges shaped their ratings according to their assumptions about a wine's price point. Several judges also explained that they assign high ratings due to the commercial value of awards and thus the long-term viability of a wine competition. Some judges assign a score to one wine according to a general zone of quality

and then score the remaining wines "around" that anchor. All of those are examples of the cognitive biases. Judges may be expressing their findings using a common scale as instructed, but their objectives and the bases for their findings may vary. A sum of judges' ratings may thus be a sum of different objectives and adherence to instructions.

(7) A sum is not an aggregate utility. Calculating sums of judges' ratings may violate precepts of economic utility theory. Barnett (2003) asserts that decision-making agents make judgments about relative preferences that are unobservable *priors* and that ratings are *post hoc* expressions of those preferences. He asserts that any implication of continuity between ratings, and the relative magnitudes of ratings, is an artifact of assumptions about a rating system rather than information about underlying utility and preferences. For example, Barnett would assert that a wine assigned a score of 10 can't be said to have twice the utility of a wine assigned a score of 5. He further asserts that such ratings cannot be employed to quantify the differences between judges' relative preferences. Marks (2019) examines wine ratings as psychophysical scaling and draws similar conclusions about the perils of comparing one person's ratings to another's. Pursuing the meaning of a sum of ratings in the context of utility theory is beyond the scope of this article. However, that issue is addressed here by affirming that ratings are reference numbers for ordered categories, and they are not interpreted here as cardinal measures of utility.

Calculating the arithmetic sums of judges' ratings on wines is common, easy to do, and easy communicate. But it's flawed as a method of assessing aggregate relative quality or preference. At least the seven factors discussed above create variance and cloud the interpretation of relative sums. If sums are to be considered, then the Turing corollary and the vinegar axiom imply that consensus about those sums ought to also be considered.

III. Sum Stars & Bars

Suppose each of three judges (each judge j and $J = 3$) assessed each of four wines (each wine i and $W = 4$) and assigned a rating of 1 through 5 ($x_{ij} \in (x_{min} = 1, \ldots, x_{max} = 5)$) to each wine. The set of ratings that the judges assign to a wine, for example, could be {1,1,3}. The total number of possible ordered sets is $(x_{max} - x_{min} + 1)^J = (5)^3 = 125$. The range of possible sums of ratings ($S_i = \sum_{j=1}^{J} x_{ij}$) is from the number of judges up to the number of judges times the highest possible rating ($S_i \in \{J, \ldots, J * x_{max}\}$). The only way to obtain the lowest or highest sum is for all judges to assign the same lowest or highest rating. But in between, the sums of various sets of ratings can yield the same sum. Calculating the number of different sets that yield a particular sum, and thus the distribution of potential sums, is an application of the binomial formula known as "Stars & Bars." See for example Earnest (2019).

Suppose we want to know how many combinations of ratings yield the sum $S_i = 4$. That sum can be symbolized as four stars (****). Inserting two bars among the stars divides them into three combinations, one for each judge, that yield $S_i = 4$ (*|*|** and *|**|* and **|*|*). Adopting terminology from Charalambides (1982, p. 133), the

collection of ratings represented by one instance of Stars & Bars is called a *composition* of S_i with J *parts*. Those three sets of Stars & Bars show that the compositions of $S_i = 4$ are {1,1,2}, {1,2,1} and {2,1,1}. Stars & Bars expresses a counting problem that can be solved using the binomial formula for "n choose k" in Equation (1). In binomial coefficient notation, a step toward calculating the number of "n multi choose k" compositions of S_i with J parts ($C_J^{S_i}$) appears in Equation (2).

$$\left(\begin{array}{c} n \\ k \end{array} \right) = \frac{n!}{k! \cdot (n-k!)} \mid n \geq k \tag{1}$$

$$C_J^{S_i} \leq \left(\begin{array}{c} S_i + J - 1 \\ J - 1 \end{array} \right) \tag{2}$$

The inequality in Equation (2) indicates that some parts of some compositions may be inadmissible. For example, one of the compositions of $S_i = 8$ is (*|*|******). Note that the last part of that composition contains six stars, but the maximum possible rating (upper bound b) is 5. Applying the inclusion exclusion theorem to exclude inadmissible parts yields the equality in Equation (3). See for example Charalambides (1982, p. 113, 136, 143) and Algorithms for Competitive Programming (2024). MATLAB code for Equations (2) and (3), and a check using a brute force enumeration of all possible compositions, is available on request. The usefulness of Equation (3) is that it provides computationally efficient solutions to large-number problems that require unrealistic computer memory and time to solve using brute force enumeration.

$$C_J^{S_i|b} = \sum_{z=0}^{\frac{S_i}{b}} (-1)^z \left(\begin{array}{c} J \\ z \end{array} \right) \left(\begin{array}{c} S_i + J - 1 - z\,(b) \\ J - 1 \end{array} \right) \tag{3}$$

Continuing the example above in which three judges assign a rating of 1 through 5 to each wine, $J = 3$ and $b = 5$. A sum of four is produced by three compositions, so if $S_i = 4$ then $C_J^{S_i|b} = 3$.

Using Equation (3), the exact distribution of the potential sums of ratings for the white wines assessed by nine judges using a 0–20 rating scale at the 1976 Judgment of Paris appears in Figure 1. The lowest possible sum is $J\,(x_{min}) = 9\,(0) = 0$, the highest possible sum is $J\,(x_{max}) = 9\,(20) = 180$, and the total number of possible ordered sets for each wine is $(x_{max} - x_{min} + 1)^J = (21)^9 = 794\ billion$.[1] The computational benefit of Equation (3) is avoiding the need to enumerate and evaluate all 794 billion possible compositions of ratings. The observed sums of ratings for each of the Judgment's 10 white wines are also indicated in Figure 1. The sums for the top three wines are within a range of 15 points, but each sum is the result of several to nearly 2 billion possible

[1]Half points were allowed in the Judgment of Paris. For simplicity, half points are omitted in this calculation of potential compositions. Including half-points increases the number of potential compositions from billions to trillions.

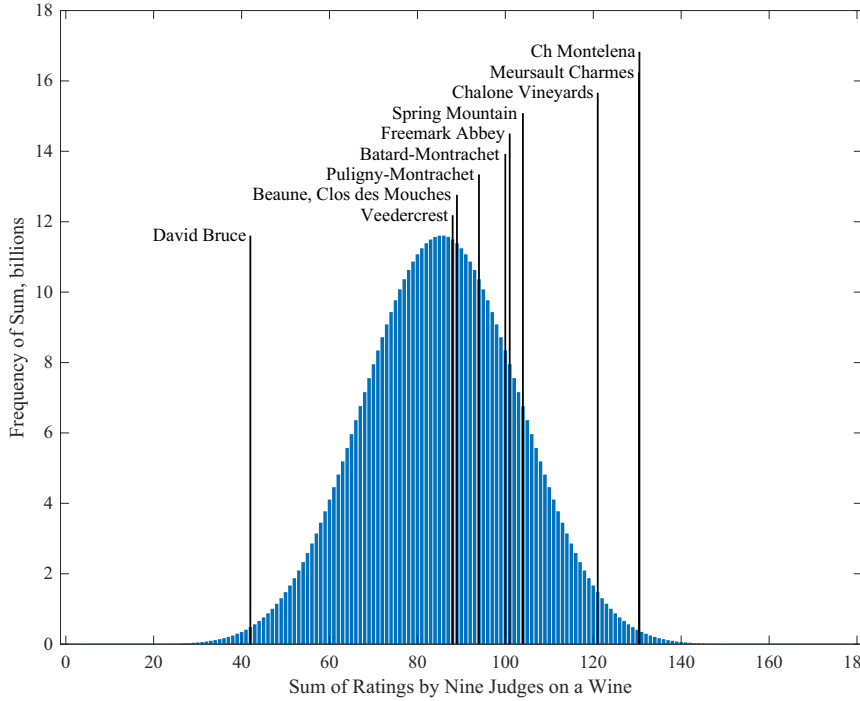

Figure 1. Are the Paris results a random illusion of consensus?

compositions of judges' scores. The wines with sums below the top three, except for lowest-sum David Bruce, are each the result of over 6–11 billion possible compositions. Considering the evidence that there is uncertainty about the scores that judges assign, Figure 1 does not inspire confidence that the observed sums indicate any more than a random illusion of consensus about what wines are better than others. What those observed scores and sums do indicate about consensus is addressed in the next section.

Before moving forward to address consensus, the results above concerned scores, but some rating systems involve ranks rather than scores. The scores considered above are ordered ratings sampled with replacement that are in theory tied to some external standard of quality. For example, every wine in a flight could earn a Gold rating if every wine met the Gold standard. In contrast, ranks are ordered ratings sampled without replacement that are in theory tied only to the relative merit of wines within a set. Ranks and points against or Borda scores can be evaluated using different parameters in Equation (3) and the results look like those in Figure 1.

IV. Sum scatter

Equation (3) and Figure 1 show that, in some cases, a sum can be the result of over billions of compositions and those compositions express differences in consensus among the judges. At maximum or perfect consensus, every judge assigns the same rating. For no consensus at all, judges' ratings are evenly distributed across the range of potential

ratings. Both of those constellations of Stars & Bars, and many in between, can have the same sum. A measure of consensus needs to be considered.

A. Indications of preference order

Cross-correlations and Kendall's coefficient of concordance provide indications of the potential for consensus about a preference order among wines in a flight. See also correlations among judges and consumers in Kopsacheilis *et al.* (Kopsacheilis et al., 2024, p. 292).

Arranging the ratings assigned by judges to the flight of white wines in the 1976 Judgment in a matrix (JxW) and then calculating the cross-correlations between judges yields the results in Table 1. Ignoring the self-correlations of 1.00, only 4 of the correlations are negative, few are near zero, and 21 of the 55 cross-correlations are greater than 0.5. Those results indicate the potential for some consensus among judges regarding a preference order.[2]

In addition, the cross-correlations in Table 1 indicate what judges, if any, don't appear to contribute to any potential for consensus. Judges #4 and #8, for example, have the most negative and lowest correlations. That finding makes sense. Ms. Gallagher and Mr. Spurier were the organizers and never planned to have their scores included with presumably more qualified judges' ratings in the official results. That finding is also an example of considering the Turing corollary presented in Section II; judges whose ratings are indistinguishable from random assignments, or uncorrelated with any other judge, should not influence determination of a non-random preference order.

Kendall (1962) proposed to measure agreement among all raters on all objects in a matrix with a single non-parametric coefficient of concordance known as Kendall's W. That statistic is calculated using ranked data and it has a range of [0, 1] where $W = 0$ means no consensus on a rank order and $W = 1$ means every judge assigned the same ranks to the same wines. Quandt (2006) employed Kendall's W to evaluate concordance among participants in the *Liquid Assets Wine Group* at Princeton University (p. 15–16). Malkiel (2024) employed W to update results for *Liquid Assets* and he found little concordance among the participants. Tinsely & Weiss (1975, p. 366) counseled against Kendall's W when judges assign scores rather than ranks because conversion of scores to ranks erases information about ratings other than their serial positions. Subject to that qualification and after converting the 1976 Judgment scores to ranks, $W = 0.43$. Like the cross-correlations in Table 1, that result indicates some but not strong consensus.

B. Dispersion in preference for a wine

While cross-correlations and Kendall's W express the potential for a preference order, neither expresses what that order may be. And when sums of ratings are used to determine that order, Section III showed that dispersion must also be considered. Equation (3) and Figure 1 showed that potentially billions of compositions, with at

[2]Taber (2005) reported that the judges discussed the wines while scoring. Some ratings assignments may thus not be independent, and that discussion may contribute to the resulting positive correlations.

Table 1. Cross-correlations for pairs of judges

Judge		1	2	3	4	5	6	7	8	9	10	11
1	P. Bréjoux	1.00										
2	A. Villaine	0.22	1.00									
3	M. Dovaz	0.65	−0.07	1.00								
4	P. Gallagher	0.70	−0.03	0.48	1.00							
5	O. Khan	0.61	0.79	0.19	0.10	1.00						
6	C. Millot	0.31	0.83	0.15	0.18	0.63	1.00					
7	R. Oliver	0.55	0.85	0.14	0.15	0.93	0.70	1.00				
8	S. Spurier	0.54	−0.07	0.18	0.44	0.09	−0.21	0.17	1.00			
9	P. Tari	0.50	0.32	0.68	0.33	0.33	0.53	0.52	0.20	1.00		
10	V. Vannequé	0.81	0.45	0.28	0.62	0.67	0.46	0.62	0.51	0.26	1.00	
11	J. Vernat	0.56	0.85	0.16	0.17	0.94	0.71	1.00	0.16	0.51	0.64	1.00

least millions of different dispersions, can have the same sum. While proposing a utility function of sum and dispersion, much like a utility function of investment rate of return and risk, is beyond the scope of this article, at least dispersion ought to be considered and measured.

Variance in the ratings that judges assign is a common measure of dispersion. Draws from a bounded set, including wine ratings, have a bounded variance. Considering the sample variance (s^2) of the small samples that are typical of wine competitions, the maximum variance (s^2_{max}) occurs at maximum disagreement when even numbers of judges cluster their assignments at only the lowest and highest ratings.[3] Tastle and Weirman (2007) proposed an alternative to variance and employed Shannon's notion of information entropy to express the entropy of dispersion among ordinal ratings such as the Likert scale. Their measure of consensus (Cns_i) is a cross-entropy function of the probability distribution of observed ratings, the absolute deviation from the mean of observed ratings, and the range of possible ratings.[4] Elzinga *et al.* (Elzinga et al., 2011, p. 2547) considered Tastle & Wierman's Cns_i and they also proposed a variance-based index, but they used large-sample population variance rather than sample variance. They express preference for C_i over Cns_i because it is easier to calculate and interpret. Following Jaynes (1957), a maximum information entropy methodology pursues the maximum entropy that is consistent with observed data and, if any, logical axioms. On that basis, consensus in a small sample should have more entropy than consensus in a large sample. That condition argues in favor of the sample-variance-based statistic C_i above.

Now we can answer the question posed in Figure 1: Are the Paris results a random illusion of consensus? Results for the sums of ratings S_i and corresponding indexes of consensus C_i appear in Figure 2. $s^2_{max} = 0.63$ So none of the sums appear to be nearly random. Chateau Montelena had the highest sum of ratings, but the second-lowest consensus and Meursault Charmes had the second-highest sum and the highest consensus.

Although considering the consensus about sums in Figure 2 does address, in part, the seven conundrums discussed in Section II, wine competition officials must still weigh sums of scores versus dispersion among scores to decide winners and losers, to decide on a rank order of quality or preference.

V. Order preference model solution

Dispersion within the sets of ordered ratings that judges assign to objects can be addressed using a probabilistic approach to estimate the most likely consensus ordering. Probabilistic order preference models have been applied to taste tests of pudding (Davidson, 1970), snap beans (Plackett, 1975), crackers (Critchlow, 1980), salad (Critchlow and Fligner, 1991), soft drinks (Bockenholt, 1992), animal feed (Marden, 1995), cheese snacks (Vigneau et al., 1999), an unidentified food

[3] $s^2_{max} = \frac{n}{n-1}\left(\frac{r_{max}-r_{min}}{2}\right)^2$

[4] $Cns_i = 1 + \sum_{j=1}^{J} \rho_j \cdot log_2\left(1 - \frac{|x_j-\mu|}{r_{max}-r_{min}}\right)$

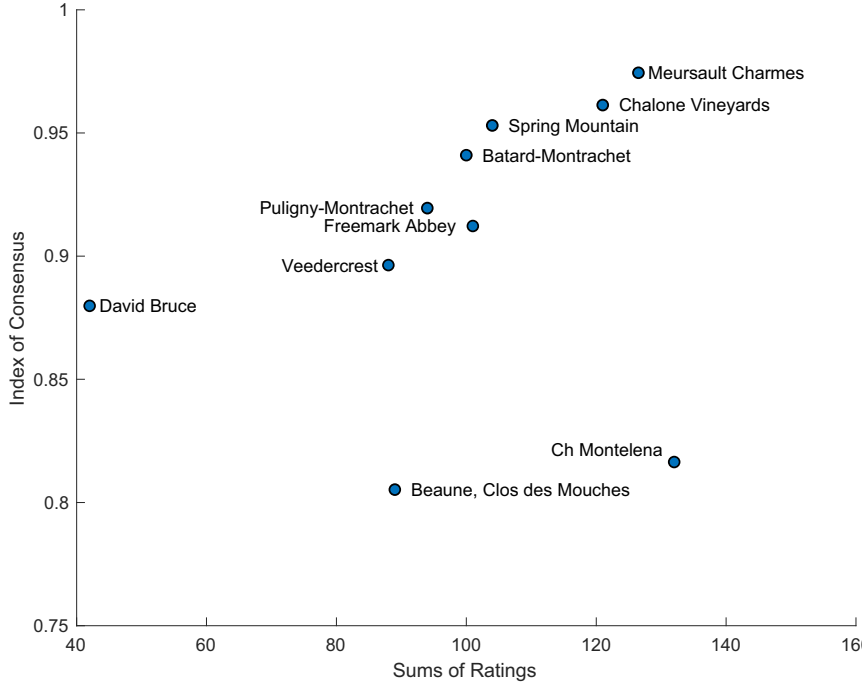

Figure 2. Was Chateau Montelena better than Meursault Charmes?

(Cleaver and Wedel, 2001), salad dressings (Theusen, 2007), sushi (Chen, 2014), sweet potato varieties (Moyo et al., 2021), and recently to wine (J.C. Bodington, 2015).

Texts by Marden (1995), and Alvo & Yu (2014) review various methods of evaluating ordered data. Both texts, and many other publications, present the Plackett–Luce model and that model was employed to evaluate the taste test results cited above by Chen, Marden, and Bodington. The general form of Plackett–Luce appears in Equation (6) below. In Equation (6a), the probability that a judge assigns a top-choice or most-preferred rating to a wine (υ_i) is employed to calculate the probability (ρ_j) of a judge's rating order vector $\left(x_j^o\right)$. The machinery in Equation (6b) can be visualized as calculating the probability of one branch on a probability tree. The log likelihood ($\mathcal{L}$) of the observed preference orders assigned by all the judges appears in (6b). Maximizing likelihood $\mathcal{L}$ using the probabilities $\hat{\upsilon}_i$ then yields the order vector probabilities ρ_i that indicate the most likely aggregate preference order for the wines. A likelihood ratio statistic (LRS) can then be employed to test the null hypothesis that the probabilities ρ_i are a random result.[5] MATLAB code for Equation (6), available on request, was checked by replicating the example in Marden (1995, p. 216).

[5]LRS $= -2\ (\mathcal{L}_{null\ hypothesis}$ - $\mathcal{L}_{solution})$ has a Chi Square distribution. The degrees of freedom (df) are the number of additional parameters in the $\mathcal{L}_{solution}$ model. For $\mathcal{L}_{null\ hypothesis}$ all $\rho_i = 1/W$. See Marden (1995, p. 143).

Table 2. Preference-order results, Paris 1976 white wines

					Plackett–Luce, Eq. (6)[b]	
Wine name	Wine #	Sums of scores[a]	Implied rank	Standard deviation	Top-choice probability	Implied rank diff from sums)
Chateau Montelena	1	130.5	1	5	0.24	2 (↓1)
Meursault Charmes	2	130.5[c]	2	1.3	0.27	1 (↑1)
Chalone Vineyard	3	121	3	2.2	0.14	3
Spring Mountain Vineyard	4	104	4	1.9	0.10	4
Clos des M. Beaune	5	101	5	4.8	0.06	6 (↓1)
Freemark Abbey Winery	6	100	6	3.2	0.07	5 (↑1)
Batard-Montrachet	7	94	7	2.4	0.05	7
Piligny-Montrachet	8	88	8	2.5	0.04	8
Veedercrest Vineyards	9	88	9	3.7	0.03	9
David Bruce	10	42	10	3	0.00	10

a) Scores taken from and results match De Nicoló (2023, p. 23).
b) LRS: 70.3 with 9 df
c) This sum of scores represents the total reported in Taber (2005). The same sum has also been reported in other sources as 126.5.

$$\rho_j \left| x_j^o = \prod_{k=1}^{W} \left(\frac{\hat{v}_i}{\sum_{k=i}^{W} (\hat{v}_k)} \right) \right| x_{kj}^o \tag{6a}$$

$$\mathcal{L}\left(x^o\right) = \sum_{j=1}^{J} log_2\left(\rho_j\right) \tag{6b}$$

Results for an application of the Plackett–Luce model to the Paris 1976 white wines appear in Table 2. The sums of scores shown in Table 2 below imply the preference order from left to right. However, the top-choice probabilities for the Plackett–Luce results imply a different order. Those results consider differences in dispersion that sums ignore and they imply a switch in the orders of wines #1 and #2 and then also in wines #5 and #6. The Plackett–Luce results also break the tie in sums for wines #8 and #9. Plackett–Luce solves the problem that top-sum Chateau Montelena had the highest sum of scores, but the highest dispersion and Meursault Charmes had the second-highest sum but the lowest dispersion.

Plackett–Luce and other preference-order models are an improvement over simple sums of ratings that ignore dispersion, but they are more difficult to employ than simple sums and they do not entirely resolve all the issues enumerated in

Section II. Preference-order models require optimization software and potentially complex adjustments are needed to reflect stochastic ratings, measure statistical significance for small sample sizes, and to express that some judges are more reliable than others.

VI. Conclusions

Sums of ratings are widely employed, easy to calculate, and easy to communicate. But (1) ratings are stochastic, (2) sample sizes are small, (3) the range effect biases influence, (4) ranks ignore differential intensity, (5) some judges are more consistent than others, (6) adherence to competition guidance is not uniform, and (7) a sum is not an aggregate utility.

Stars & Bars combinatorics shows that many billions of compositions of ratings can have the same sum and that uncertain ratings can have many different sums. Dispersion among the ratings that compose a sum can range from little to none when every judge assigns the same rating, to a uniform distribution, to extreme disagreement where judges' assignments cluster at the highest and lowest ratings. If sums of ratings are to be considered, that dispersion must also be but is difficult to consider.

A Plackett–Luce preference order model avoids the difficulty of considering both sums of scores and dispersion. That model yields a maximum likelihood estimate of the top-choice probability for each wine, and those probabilities yield a defensible order of quality or preference. And an application to the white wines tasted at the 1976 Judgement of Paris implies a different preference order than the order implied by sums of ratings. For example, the maximum likelihood is that the first-place wine was French Meursault Charmes rather than Californian Chateau Montelena.

Application of order preference models to wine competition results is nascent. Applying the Plackett–Luce and other models to the 1976 Paris and other data could be considered. Further research concerning methods of preserving the information in scores, improving a top-choice probability density function, discounting the influence of judges with low cross-correlations, refining maximum entropy considerations, and addressing the conundrums enumerated in Section II seems worthwhile.

Acknowledgements. The author thanks an anonymous reviewer and Neal Hulkower, PhD, for their essential encouragement, ideas, and comments.

References

Algorithms for Competitive Programming (2024). The inclusion-exclusion principle. Available from https://cp-algorithms.com/combinatorics/inclusion-exclusion.html#number-of-upper-bound-integer-sums (accessed 5 August 2024).

Alvo, M., and Yu, P. L. H. (2014). *Statistical Methods for Ranking Data* (273). Springer.

Ashenfelter, O., and Quandt, R. E. (1999). Analyzing a wine tasting statistically. *Chance*, *12*(3), 16–20. https://doi.org/10.1080/09332480.1999.10542152

Barnett, W. (2003). The modern theory of consumer behavior: Ordinal or cardinal? *Quarterly Journal of Austrian Economics*, *6*(1), 41–65. https://doi.org/10.1007/s12113-003-1012-4

Bockenholt, U. (1992). Thurstonian representation for partial ranking data. *British Journal of Mathematical and Statistical Psychology*, *45*(1), 31–49. https://doi.org/10.1111/j.2044-8317.1992.tb00976.x

Bodington, J. (2022). Stochastic error and biases remain in blind ratings. *Journal of Wine Economics*, *17*(4), 345–351. https://doi.org/10.1017/jwe.2022.53

Bodington, J. C. (2015). Testing a mixture of rank preference models on judges' scores in Paris and Princeton. *Journal of Wine Economics*, *10*(2), 173–189. https://doi.org/10.1017/jwe.2015.18

Charalambides, C. A. (1982). On the enumeration of certain compositions and related sequences of numbers. *The Fibonacci Quarterly*, *20*(2), 132–146. https://doi.org/10.1080/00150517.1982.12430010

Chen, W. (2014). How to Order Sushi. PhD Dissertation, Harvard University.

Circle of Wine Writers (2025). Let' talk about wine judging, practice and purpose. Available from https://www.youtube.com/watch?v=ET3SsuWMIqU. April 1, 2025 (accessed April 14, 2025).

Cleaver, G., and Wedel, M. (2001). Identifying random-scoring respondent in sensory research using finite mixture regression results. *Food Quality and Preference*, *12*(5–7), 373–384. https://doi.org/10.1016/S0950-3293(01)00028-3

Cochran, W. G. (1937). Problems arising in the analysis of a series of similar experiments. *Supplement to the Journal of the Royal Statistical Society*, *4*, 102–118. https://doi.org/10.2307/2984123

Critchlow, D. E. (1980). *Metric Methods for Analyzing Partially Ranked Data*. Springer-Verlag.

Critchlow, D. E., and Fligner, M. (1991). Paired comparison, triple comparison, and ranking experiments as generalized linear models, and their implementation on GLIM. *Psychometrika*, *56*(3), 517–533. https://doi.org/10.1007/BF02294488

Davidson, R. R. (1970). On extending the Bradley-Terry model to accommodate ties in paired comparison experiments. *Journal of the American Statistical Association*, *65*(329), 317–328. https://doi.org/10.1080/01621459.1970.10481082

De Nicoló, G. (2025). Wine ratings and commercial reality. *Journal of Wine Economics*, *20*(1), 1–25. https://doi.org/10.1017/jwe.2024.27

Earnest, M. (2019). Mathematics, extended star-and-bars (where the upper limit of the variable is bounded). Available from https://math.stackexchange.com/questions/553960/extended-stars-and-bars-problemwhere-the-upper-limit-of-the-variable-is-bounded. 10 April 2019 (accessed 16 July 2024).

Elzinga, C., Eang, H., Lin, Z., and Kumar, L. (2011). Concordance and consensus. *Information Sciences*, *182*(12), 2529–2549. https://doi.org/10.1016/j.ins.2011.02.001

Jaynes, E. T. (1957). Information theory and statistical mechanics. *Physical Review*, *106*(4), 620–630. https://doi.org/10.1103/PhysRev.106.620

Kahneman, D., Sibony, O., and Sunstein, C. R. (2021). *Noise, a Flaw in Human Judgement*. Little, Brown Spark, Hachette Book Group, 454.

Kendall, M. G. (1962). *Rank Correlation Methods*. Hafner Publishing Company, 199.

Kopsacheilis, O., Analytis, P. P., Kaushik, K., Herzog, S. M., Bashrami, B., and Deroy, O. (2024). Crowdsourcing the assessment of wine quality: Vivino ratings, professional critics, and the weather. *Journal of Wine Economics*, *19*(3), 285–304. https://doi.org/10.1017/jwe.2024.20

Malkiel, B. (2024). Tastings at tea time: The Princeton wine group. *Journal of Wine Economics*, *19*(2), 99–112. https://doi.org/10.1017/jwe.2024.5

Marden, J. I. (1995). *Analyzing and Modeling Rank Data*. Chapman & Hall, London, 329.

Marks, D. (2019). If this wine got 96 out of 100 points, what is wrong with me? A critique of wine ratings as psychophysical scaling. *American Association of Wine Economists, Working Paper No*, *239*, 24. https://doi.org/10.1017/jwe.2020.42

Moyo, M., Ssali, R., Namanda, S., Nakitto, M., Dery, E. K., Akansake, D., Adjebeng-Danquah, J., van Etten, J., de Sousa, K., Lindqvist-Kreuze, H., Carey, E., and Muzhingi, T. (2021). Consumer preference testing of boiled sweet potato using crowdsourced citizen science in Ghana and Uganda. *Frontiers in Sustainable Food Systems*, *5*, 620363. https://doi.org/10.3389/fsufs.2021.620363

OIV (2021). *OIV Standard for International Wine and Spirituous Beverages of Viticultural Origin Competitions*. 2021. International Organization of Vine and Wine, 25.

Plackett, R. L. (1975). The analysis of permutations. *Applied Statistics*, *24*(2), 193–202. https://doi.org/10.2307/2346567

Quandt, R. E. (2006). Measurement and inference in wine tasting. *Journal of Wine Economics*, *1*(1), 7–30. https://doi.org/10.1017/S1931436100001826

Taber, G. M. (2005). *Judgment of Paris: California Vs. France and the Historic 1976 Paris Tasting that Revolutionized Wine*. Scribner, 327.

Tastle, W. J., and Weirman, M. J. (2007). Consensus and dissention: A measure of ordinal dispersion. *International Journal of Approximate Reasoning*, *24*(3), 531–545. https://doi.org/10.1016/j.ijar.2006.06.024

Theusen, K. F. (2007). Analysis of Ranked Preference Data. Informatics and Mathematical Modeling, Masters Thesis. Technical University of Denmark, Kongens Lyngby.
Tinsely & Weiss. (1975). Interrater reliability and agreement of subjective judgements. *Journal of Counseling Psychology*, *22*(4), 358–376. https://doi.org/10.1037/h0076640
Turing, A. M. (1950). Computing machinery and intelligence. *Mind, New Series*, *59*(236), 433–460. https://doi.org/10.1093/mind/LIX.236.433
Vigneau, E., Courcoux, P., and Semenou, M. (1999). Analysis of ranked preference data using latent class models. *Food Quality and Preference*, *10*(1999), 201–207. https://doi.org/10.1016/S0950-3293(99)00017-8

Cite this article: Bodington, J. (2025). Wine Stars & Bars: The combinatorics of critic consensus and the usefulness of order preference models. *Journal of Wine Economics*, 1–14. https://doi.org/10.1017/jwe.2025.10093