Journal of Wine Economics, Volume 12, Number 3, 2017, Pages 267–281
doi:10.1017/jwe.2017.21

Disentangling Wine Judges' Consensus, Idiosyncratic, and Random Expressions of Quality or Preference*

Jeffrey C. Bodington [a]

Abstract

Judges confer various awards on wines entered in dozens of wine competitions each year. This article employs data on blind replicates to show that those awards are based on one instance of stochastic ratings assigned by wine judges; awards based on the expected values of those stochastic ratings would be different. This article recognizes the stochastic nature of ratings and builds on the work of many others to propose and test a conditional-probability model that yields maximum-likelihood estimates of judges' latent consensus, idiosyncratic, and random assignments of scores to wines. The exact p-value for a likelihood test of the null hypothesis that the model's results are random is less than 0.001. Applying the notion of conditional probability may lead to better methods of assigning awards to entries in wine competitions and of assessing the capabilities of wine judges. (JEL Classifications: A10, C10, C00, C12, D12)

Keywords: consensus, idiosyncratic, random, statistics, wine tasting.

I. Introduction

Judges confer medals, ribbons, scores, ranks, and other awards on wines entered in dozens of wine competitions each year. Diverse literature implies that those awards are the observable results of an unseen or latent mixture of judges' consensus, idiosyncratic, and random decisions about quality or preference. Further, experiments with blind replicates in wine competitions show that the random component of judges' decisions is material, variable, and nuanced.

The awards noted above are usually conferred using the sums of scores or the sums of ranks assigned by a small number of judges. Those methods are easy to use and communicate. Some competitions and researchers use or are examining Borda counts, Shapely values, and preference models (see also Cao and Stokes, 2017;

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

Ginsburgh and Zhang, 2012). Regardless of which of those five methods is employed, it yields an aggregation that is based on judges' observed ratings but ignores the latent randomness that is a foundation of those ratings. Considering that foundation would often lead to different awards. Disentangling the latent consensus, idiosyncratic, and random components of judges' ratings can yield awards that are closer to a mode, mean, or maximum likelihood of consensus. Disentangling the latent components of judges' ratings can also yield useful information about the judges and the wines.

Section II begins with analysis of the distribution of the ratings assigned to blind replicates. The results are then employed in Section III to show that aggregations of ratings are conditional and unlikely to yield a mode, mean, or maximum likelihood of consensus. Then, building on the literature and work of others, a model is proposed and tested in Section IV that uncovers the latent consensus, idiosyncratic, and random components of judges' ratings. Using the Stellenbosch data published in Cicchetti (2014) as an example, the exact p-value for the null hypothesis that the model obtained a random result is <0.001. Conclusions and discussion follow in Section V.

Before moving forward, the author's anecdotal experience is that many Master Sommeliers, Masters of Wine, Wine and Spirit Education Trust (WSET) certificate holders, and other wine professionals express disdain for scoring wines and quantitative analyses of those scores.[1] They assert that wines and tasters are too complex and too idiosyncratic for scores to convey much useful information. Nevertheless, every year, and most often judged by wine professionals, dozens of state fair, county fair, magazine, newspaper, and other wine competitions and reviews confer ribbons, medals, awards, ranks, and scores. All of those designations can be expressed as ranks or scores.[2] This article is an effort to analyze the designations awarded to wines while keeping complexity and idiosyncrasy in sight.

II. The Probability Distribution of an Observed Rating

Hodgson (2008), Ashton (2012), Hodgson and Cao (2014), and Cicchetti (2014) show that a wine judge with near-perfect consistency, one who assigns the same rating to the same wine every time, is rare. Bodington (2017b) finds that wine judges tend to assign closer ratings to replicates than is likely due to chance alone.

[1] Disclosure: the author holds WSET Level II and Level III certifications.

[2] For example, the WSET (2014, 3) systematic approach concludes by designating a wine as faulty, poor, acceptable, good, very good, or outstanding. That is an ordered set with six ranks; sampling is with replacement. Liquid Assets tasters assign ranks in order of relative preference; sampling is without replacement (liquidasset.com). Among several methods of transforming scores into ranks, Ashton (2016, 267) expresses Robert Parker's quality-level scores as a set with eight ranks and Jancis Robinson's scores as a set with eleven ranks. In both cases, sampling is with replacement. As shown in Olkin, Lou, Stokes, and Cao (2015, 9) and Bodington (2015b, 175, 179), ties between wines are merely the expectations of rank permutations. Finally, ranks can be transformed into evenly spaced scores on any scale.

He also concludes that the distribution of ratings assigned to blind replicates is determined by judges' capabilities, the mechanics of the tasting protocol, and the difference between the replicate and the other wines in the flight.

The findings summarized above are expressed by the probability mass function (PMF) in Equation (1). The probability of an observed score (f) for a particular replicate wine (i, and a total of W wines) for a particular judge (j, and a total of J judges) is an exponential function of the observed score ($s_{j,i}$), a modal score parameter ($\hat{s}_{j,im}$), the standard deviation of the judge's scores on all the wines in a tasting (σ_j), and a dispersion parameter ($0 \leq \hat{\theta}_j \leq 1$). The PMF in Equation (1) expresses a discrete, unimodal, and bounded distribution. For $\hat{\theta} = 0$ and $s_{j,i} = \hat{s}_{j,im}$, the probability of $s_{j,i}$ is unity. That is perfect consistency, meaning that a judge assigns the same score to the same wine every time. For $\hat{\theta} = 1$, the PMF describes a distribution in which the probability of every $s_{j,i}$ is the same. In that case, there is no consistency, and a judge assigns scores as if they were drawn from a uniform random distribution:

$$f\left(s_{j,i}|\hat{\theta}_j,\ \hat{s}_{j,im}\right) = \left(\frac{1}{C_j}\right)\hat{\theta}_j^{d_{j,i}} \tag{1A}$$

$$d_{j,i} = \left(\frac{s_{j,i} - \hat{s}_{j,im}}{\sigma_j}\right)^2 \tag{1B}$$

$$C_j = \sum\nolimits_{s_{min}}^{s_{max}} \hat{\theta}_j^{d_{j,i}} \tag{1C}$$

The distance (d) defined in Equation (1B) is the square of the standardized difference between the observed and modal scores. First, considering the distance ($s_{j,i} - \hat{s}_{j,im}$) alone can lead to a mirage of consistency. Some judges spread their scores more broadly over the allowed range than others. For example, at Stellenbosch, $\sigma_7 = 1.6$ and $\sigma_5 = 11.4$. Thus, a judge with a narrow spread on replicates can appear to be highly consistent even if all of his or her scores are also assigned randomly within a narrow range. Dividing distance by σ_j standardizes the difference and then favors judges who assign scores to replicates within a narrower range than the scores that each judge assigns to all the wines. An additional benefit of standardizing is that distance becomes unit-less, so $\hat{\theta}_j$ can be compared across tastings that have difference score ranges. Finally, the constant (C_j) in Equation (1C) normalizes the results of the exponential function in Equation (1A) so that the sum of probabilities equals unity.

A test and example of Equation (1) appears in Figure (1). Stellenbosch Judge #7 assigns scores of (83, 84, 85) to replicates of Sauvignon Blanc and $\hat{\theta}_7 = 0.13$, Judge #3 assigns (78, 82, 85) and $\hat{\theta}_3 = 0.14$, and Judge #4 assigns (65, 72, 86) and $\hat{\theta}_4 = 0.74$. The maximum likelihood estimates (MLEs) of the parameters in Equation (1) for all 15 judges appear in Table 1.

Table 1

MLEs of Parameters in Equation (1) for Triplicates of Stellenbosch, Sauvignon Blanc

	Judge														
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*
$\hat{s}_{j,im}$	81.3	75.0	81.7	74.3	75.6	77.0	84.0	83.3	86.7	94.7	85.3	62.7	80.3	79.7	82.7
$\hat{\theta}_j$	0.24	0.39	0.14	0.74	0.74	0.50	0.13	0.60	0.54	0.50	0.67	0.32	0.04	0.09	0.31

III. Start Over

The results in Figure 1 show that the scores that judges assign to a wine are not identically distributed. The results in Figure 1 for Judge #4 show that one draw from the distribution of scores for that judge is unlikely to be even close to the expected value of scores. More important, the result of a function with a stochastic input is also stochastic. Sums of scores, sums of ranks, Shapely values, Borda counts, and preference-model results are therefore conditional, and they depend on one instance of stochastic ratings. Especially for the small sample sizes that are typical of wine competitions, the relationship between results for one instance of ratings and the mode, mean, or maximum likelihood of results for the potential range of instances is thus unknown. Without further investigation, little can be said about the interpretation and reliability of conditional results that are aggregated while ignoring the implications of Figure 1. The appearance of a reliable consensus within observed ratings, whether based on sums of scores, Borda, or any other metric, may be an illusion.

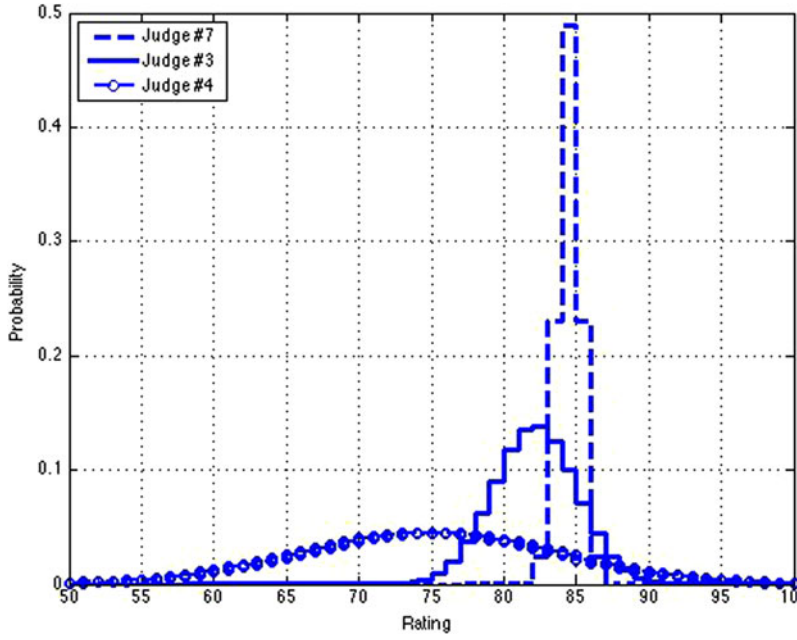
As a test and example, based on simple sums of scores, the aggregate order of the quality ratings assigned by 15 judges to the Stellenbosch Sauvignon Blanc is (8 T, 5 T, 6, 1, 7, 3, 4, 2 T). Wine #1 is ranked fourth, #2 T is ranked last, and wine #8 T is ranked highest. "T" indicates that the wine is a member of the blind triplicate; thus, the sums of scores imply that the same wine from the same bottle ranks as highest and lowest quality. Using the PMF in Equation (1), the expected value of a judge's score on a wine appears in Equation (2). According to the expected values of the sums of scores, the order of quality rating for the flight of Sauvignon Blanc is (6, 2 T, 5 T, 8 T, 1, 7, 3, 4). Note that the triplicate wines now correctly group together. MATLAB code written by the author for those results is available on request. In concept, the reason for the change in order is that low-randomness judges (such as Judge #7) have more influence on differences between expected values than high-randomness judges (such as Judge #4). That effect also applies to the orders implied by sums of ranks, Shapley values, Borda counts, and preference-model results:

$$E\left(s_{j,i}\right) = \sum_{k=s_{min}}^{s_{max}} k \cdot f\left(k|\hat{\theta}_j,\ \hat{s}_{j,im}\right) \tag{2}$$

Although the example above shows that the order implied by the expected values of ratings is not the same as the order implied by the observed instance of ratings, it also shows the difficulty of analyzing wines without replicates. The order of the wines without replicates (6, 1, 7, 3, 4) does not change, because, with only one rating by each judge, there are not enough data to support PMFs for those wines. Even with blind triplicates, 3 points are meager support for estimates of $\hat{s}_{j,im}$ and $\hat{\theta}_j$. Bodington (2015a, 2015b) addresses that difficulty by parsing observed ratings in a mixture model with nonrandom and random components, and the PMF for the random component is parameterized *a priori* as a uniform random distribution. Cicchetti (2017) tests the hypothesis that judges who assign consistent scores to

Figure 1

Equation (1) PMF for Scores Assigned to Stellenbosch Replicates of Sauvignon Blanc



replicates also assign scores to nonreplicates that are consistent with the consensus of scores assigned to those wines by other judges. This article aims in the next section to build on that work and to recognize that although all judges' scores are stochastic to some extent, few judges assign scores as if drawn from a uniform random distribution, most do better, and some are as accurate as Judge #7.

IV. Disentangling Consensus, Idiosyncratic, and Random Ratings

This section moves forward in three steps. It begins with a review of consumer-choice literature and efforts to disentangle consumers' consensus and idiosyncratic preferences. That review provides useful background and definitions of consensus and idiosyncrasy, but it also shows that often-employed utility models have limited application to wine-tasting results. Second, this section presents a review of preference models, showing that preference models have been employed to evaluate the results of taste tests since the 1970s and that such models have easy application to wine-tasting results. Building on that work and Sections II and III, the third step proposes and tests a model of judges' aggregate consensus, idiosyncratic, and random wine ratings.

A. Consumer Choice, Consensus, and Idiosyncrasy

The literature on consumer choice and heterogeneity is wide and deep. Greene and Hensher (2010), Keane and Wasi (2013), Train (2002), and Yue et al. (2015) provide recent reviews of the methods and literature.

Many published evaluations of consumer choice employ utility theory, and many of those express utility in the general form $U_{j,i} = c_i + \beta_i\ A_{j,i} + \varepsilon_{j,i}$, where the utility of product i to consumer j ($U_{j,i}$) is a product-specific intrinsic utility (c_i), plus the utility due to a vector of unit values (β_i) multiplied by a vector of observable product and consumer attributes ($A_{j,i}$), plus a product- and consumer-specific idiosyncratic utility ($\varepsilon_{j,i}$). In application to wine-tasting results, research to date indicates that no covariates $A_{j,i}$ are observable and reliable predictors of judges' scores or rank assignments. In particular, Frost and Nobel (2002) review the literature, quantify the wine knowledge and sensory expertise of 57 tasters, and then obtain those tasters' hedonic ratings on 14 sensory properties and their preferences for 12 red wines. They conclude that, with the possible exception of preferences for vanilla/oak and against leather/sour flavors, expressions of preference "could not be modeled well from the sensory properties" (283). Rather than employing hedonic assessments of sensory properties, Cortez, Cerdeira, Almeida, Matos, and Reis (2009) and Nachev and Hogan (2013) employ 11 laboratory-determined physiochemical properties of wine and machine-learning methods to predict the mean scores assigned by experienced wine judges. For the one tasting that they both evaluate, they obtain accuracies up to approximately ±20%. Further research appears necessary to prove the broad application and to improve the accuracy of such analysis, and extensive physiochemical data are rarely available. Frost and Nobel also find that sensory expertise and wine knowledge are independently distributed and that no significant differences in preference "[are] found across groups based on performance in the wine knowledge test or overall expertise" (2002, 284). Mantonakis, Rodero, Lesschaeve, and Hastie (2009, 1311) find that "high knowledge" wine tasters are more prone than "low knowledge" tasters to primacy and recency biases. Ashton (2014) compares the scores assigned by novices and wine professionals to wines from California and New Jersey. He finds that the results "do not support the idea that professionals and novices differ in their appreciation for New Jersey vs. California wines" (310). Bodington (2017a) shows that female and male tasters assign about the same scores and ranks to the same wines. In sum, that literature implies that no observable attributes of either wines or judges are good predictors of the rating that a judge assigns to a wine. Consequently, in application to wine-tasting results and until future research uncovers useful covariates and methods, $U_{j,i} = c_i + \beta_i\ A_{j,i} + \varepsilon_{j,i}$ reduces to $U_{j,i} = c_i + \varepsilon_{j,i}$.

The intercept c_i is the intrinsic utility of a product and, all other things equal, reflects consumers' consensus about preference for or the quality of the product under consideration. For $c_i > c_k$, product i is preferred to or is higher quality than product k. The literature on idiosyncratic utility $\varepsilon_{t,i}$ offers many examples. Bayer, Ferreira, and McMillan (2003) examine a real-estate market, including the idiosyncratic utility of home i to buyer k. Mc Breen, Goffette-Nagot, and Jensen (2009)

evaluate a market for rental housing and find that "idiosyncratic tastes give some monopoly power" to landlords (p. 5). Hastings, Kane, and Staiger (2006) analyze school choice, including the "idiosyncratic preference of a student" for school i (p. 11). Rhee, de Palma, and Thisse (1998) evaluate the so-called first-mover advantage, consumers' "unobservable … idiosyncratic preferences," and find that being a first mover is a disadvantage when consumers have sufficiently large idiosyncratic preferences (p. 15). Rajan and Sinha (2008) evaluate hypothetical product pricing in a duopoly and find an inverse relationship between price competition and the strength of "idiosyncratic" reactions to the good (p. 3). All of those authors define and model idiosyncratic preference and quality ratings as having a distribution around c_i. Mc Breen et al. (2009) assume that $\varepsilon_{j,i}$ has a normal distribution. Bayer et al. (2003) and Hastings et al. (2006) assume that $\varepsilon_{j,i}$ has an extreme value distribution. Rajan and Sinha (2008) assume that $\varepsilon_{j,i}$ has a double exponential distribution, and Rhee et al. (1998) treat the difference between idiosyncratic preferences for two products as a random disturbance with a logistic distribution.

Although the notions of latent consensus c_i and idiosyncratic preferences $\varepsilon_{j,i}$ cited above do apply to wine-tasting results, the methodologies do not. All of those methods rely on latent utility and functions that are continuous and unbounded. In a wine tasting, judges' scores and ranks are explicit and observable measures of utility. Judges assign scores from a bounded line, and they assign ranks from a discrete, ordered, and bounded set. When ties are not allowed, sampling is without replacement. There is no support for assuming that idiosyncratic assignments have a normal, extreme-value, double-exponential, or logistic distribution. Although examining consensus and idiosyncratic ratings is common, a different methodology is necessary for examining the results of wine tastings.

B. Rank-Preference Model Applications to Taste Tests

Wine tastings, and many other applications, involve a set of objects that can be expressed as an object vector $o = (o_A, o_B, o_C, \ldots)$. Judges consider the objects and then assign to each a rating that is an assessment of absolute quality or relative preference. Those ratings can be expressed, for each judge, as a score $s_j = (s_A, s_B, s_C, \ldots)$ vector and/or a rank $r_j = (r_A, r_B, r_C, \ldots)$ vector. So-called rank-preference models are employed to examine the relationships, such as the consensus about order of quality or preference, between the vectors of judges' ratings.[3] In contrast to the linear-utility models summarized in Section IV.A, rank-preference models can be tailored to discrete, ordered, and bounded ratings that are assigned with or without replacement. The works of Marden (1995) and Alvo and Yu (2014) are widely cited texts concerning these models.

[3] The term *rank-preference model* in this article refers to the set of models that can be employed to examine scores or ranks that are indications of absolute quality or relative preference. The mechanics of a tasting protocol determine which of many potential models can and ought to be employed.

Rank-preference models have been applied to taste tests of breakfast foods (Green and Rao, 1972), snap beans (Plackett, 1975), crackers (Critchlow, 1980), soft drinks (Bockenholt, 1992), animal feed (Marden, 1995), cheese snacks (Vigneau, Courcoux, and Semenou, 1999), salad dressings (Vargo, 1989; Theusen, 2007), an unidentified food (Cleaver and Wedel, 2001), sushi (Chen, 2014), and, recently, wine (Bodington, 2015a, 2015b, 2017a). A generalized Mallows (1957) preference model is proposed in Equation (3), because it employs scores and is a simple variation of the exponential PMF already explained in Equation (1). In Equation (3), the probability of one judge's score vector (f') is the product of the probabilities that the judge assigns each score to each wine in that vector (f'^{i}). With two important exceptions, f'^{i} on the right-hand side is defined the same as it is in Equations (1A) through (1C). The exceptions are that the parameter $\hat{s}_{ic}$ is the judges' consensus score for the subject wine, and the parameter $\hat{\theta}_i$ expresses dispersion about that consensus due to idiosyncratic assignments of scores,

$$f'(s_j) = \prod_{i=1}^{W} f'^{i}\left(s_{j,i}|\hat{\theta}_i,\ \hat{s}_{ic}\right) \tag{3}$$

This article has now defined two PMFs. Equation (1) is a PMF for the probability that a judge assigns a particular score to a particular wine. Assuming a vector of such scores for every judge, Equation (3) is then a PMF for the probability of one judge's score vector within the distribution of all the judges' score vectors. Those PMFs are combined into a conditional-probability model of consensus, idiosyncratic, and random assignments below.

C. Consensus, Idiosyncrasy, and Randomness

A likelihood function that expresses the aggregate of judges' latent consensus, idiosyncratic, and random assignments of scores appears in Equation (4). The log likelihood ($\mathcal{L}$) of the observed scores is the log sum of the probability of each judge's score on each wine $f'^{i}\left(s_{j,i}|\hat{\theta}_i,\ \hat{s}_{ic}\right)$ multiplied by the probability of observing that score $f\left(s_{j,i}|\hat{\theta}_j, \hat{s}_{j,im}\right)$. Equations (1) and (3) are combined in Equation (4) to express a conditional probability. MLEs of $\hat{s}_{ic}$ yield the judges' consensus scores, MLEs of $\hat{\theta}_i$ yield the dispersion in judges' scores due to idiosyncratic differences between judges, and MLEs of $\hat{\theta}_j$ yield the dispersion in each judge's scores due to individual underlying randomness:

$$\mathcal{L} = \sum_{j=1}^{J} \sum_{i=1}^{W} ln\left(f'^{i}\left(s_{j,i}|\hat{\theta}_{ic},\ \hat{s}_{ic}\right) \cdot f\left(s_{j,i}|\hat{\theta}_j, \hat{s}_{j,im}\right)\right) \tag{4}$$

Section III concludes by noting that, with only one rating from each judge, there are not enough data to support estimates of the parameters in $f\left(s_{j,i}|\hat{\theta}_j, \hat{s}_{j,im}\right)$ for wines without replicates. Even when there are blind triplicates, again, three observations

provide meager support for estimates of two parameters. The solution proposed below employs all of a judge's scores to estimate $\hat{\theta}_j$.

Cicchetti (2017) hypothesizes that "those wine tasters who agreed reliably with their own previous evaluations of the same wine would also agree reliably with other tasters. Conversely, those tasters who disagreed with their previous evaluations of the same wines would also disagree substantially with the evaluations of other tasters." Here, that idea is restated as the hypothesis that $\hat{\theta}_j$, the underlying randomness in a judge's ratings, is positively correlated with the difference between a judge's observed scores and the all-judge-aggregate consensus scores $\left(s_{j,i} - \hat{s}_{ic}\right)$ on all wines. The PMF $f\left(s_{j,i}|\hat{\theta}_j, \hat{s}_{j,im}\right)$ in Equation (4) is thus restated here as $f\left(s_{j,i}|\hat{\theta}_j, \hat{s}_{ic}\right)$. This approach uses all the data, preserves degrees of freedom, and yields an estimate of $\hat{\theta}_j$ for each judge.

Results for the Sauvignon Blanc data appear in Table 2. Focusing on Judge #7, who has the narrowest distribution of scores in Figure 1, Table 2A shows that and $\hat{\theta}_7 = 0.31$. That finding is consistent with the stand-alone analysis of triplicates in Section II, Judge #7 is among the most accurate judges. Similar findings apply to Judge #4. Judge #4 has the broadest distribution of scores in Figure 1, $\hat{\theta}_4 = 0.77$ in Table 2A, and Judge #4 is among the less-accurate judges. The order of quality implied by the consensus scores $\hat{s}_{ic}$ in Table 2B is (7, 8 T, 5 T, 6, 2 T, 3, 4, 1). Note that the triplicates nearly group together even though no information in Equation (4) identifies them as the same wine.

If judges assign ratings as if they are drawn from a uniform random distribution, the asymptotic log likelihood according to Equation (4) is $\mathcal{L} = J \cdot W \cdot ln(1/S \cdot 1/S)$ and $15 \cdot 8 \cdot ln(1/51 \cdot 1/51) = -943.6$. MLEs of parameters shown in Table 2 for the Sauvignon Blanc data yield $\mathcal{L} = -803.5$. A chi-square test of the likelihood-ratio test statistic for the null hypothesis that those two likelihoods are the same has a p-value < 0.001. However, Pearson (1900, 166) recommends using what is now called an exact distribution if the chi-square distribution "is a bad fit" to the exact distribution. That is a risk with the small sample sizes that are typical of wine tastings. An exact random distribution of $\mathcal{L}$ is calculated using 1,000 sets of scores drawn from a uniform random distribution. Using that distribution, the exact p-value for a test of the null hypothesis that the findings above are a random result is also < 0.001.

Finally, are wine judges consistent in their inconsistency? Cicchetti (2017) concludes that replicates are "moderately confirmative" predictors of consistency in nonreplicate scores for one flight but "minimally confirmative" for another. That question is answered here by comparing estimates of dispersion in replicates alone using Equation (1) and $\hat{\theta}_j$ in Table 1 to estimates of aggregate dispersion due to randomness using Equation (4) and $\hat{\theta}_j$ in Table 2B. A scatterplot of the result appears in Figure 2. Although the slope of a least-squares line through the scatter is 0.34, the R^2 is 0.29. The correlation coefficient is 0.54. At minimum, in agreement with Cicchetti,

Table 2A
MLEs of Judge-Related Parameters in Equation (4)

	Judge														
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*
$\hat{\theta}_j$	0.56	0.73	0.51	0.77	0.86	0.82	0.31	0.59	0.70	0.94	0.64	1.00	0.80	0.80	0.62

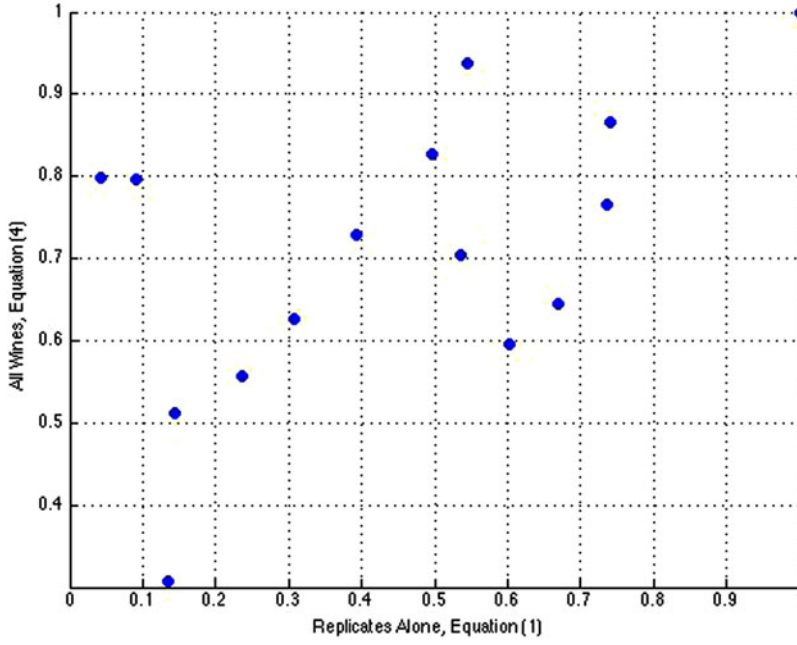
Table 2B

MLEs of Wine-Related Parameters in Equation (4)

	Wine, T indicates a member of the triplicate							
	1	*2 T*	*3*	*4*	*5 T*	*6*	*7*	*8 T*
Consensus score, $\hat{s}_{ic}$	80.0	81.5	81.1	81.0	83.3	82.9	84.1	83.9
Idiosyncratic dispersion, $\hat{\theta}_i$	0.50	0.78	0.61	0.69	0.32	0.35	0.85	0.72

Figure 2

MLEs of $\hat{\theta}_j$ Based on Replicates Alone (Equation (1)) and All Wines (Equation (4)).



the results show that dispersion in a judge's scores on replicates may not have robust implications about the consistency of scores on other wines.

V. Conclusion and Discussion

Judges confer medals, ribbons, scores, and other awards on wines entered in dozens of wine competitions each year. Section II shows that those ratings are usually more accurate than entirely random, yet still stochastic. Section III shows that sums of scores, sums of ranks, Borda count, Shapley value, and preference-model results

are conditional results. Using the notion of a conditional probability, a model is proposed and tested in Section IV that yields information about judges' latent consensus, idiosyncratic, and random expressions of quality or preference. Using data for a tasting of eight Sauvignon Blanc wines that contain a blind triplicate, the conditional-probability model detects the similarity between the triplicates, and the model results also show that the scores that a judge assigns to replicates may not be a robust guide to the accuracy of the scores that the judge assigns to other wines.

These findings are based on one model and one set of data. Tests of other models and tests using other data appear worthwhile. Other models could have different PMFs. In particular, methods of estimating PMFs that express the stochastic nature of the scores that judges assign need to be improved. The model proposed above applies to scores assigned with replacement, but another model could be developed for application to ranks assigned without replacement. Tests using other data would illuminate the general applicability and usefulness of the proposed and other models. The results may lead to more robust methods of assigning awards to entries in wine competitions and to better methods of assessing the capabilities of wine judges.

References

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

Ashton, R. H. (2012). Reliability and consensus of experienced wine judges: Expertise within and between? *Journal of Wine Economics*, 7(1), 70–87.

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.

Ashton, R. H. (2016). The value of expert opinion in the pricing of Bordeaux wine futures. *Journal of Wine Economics*, 11(2), 261–288.

Bayer, P., Ferreira, F., and McMillan, R. (2003). *A Unified Framework for Measuring Preferences for Schools and Neighborhoods*. Yale University, Economic Growth Center, Center Discussion Paper No. 872.

Bockenholt, U. (1992). Thurstonian representation for partial ranking data. *British Journal of Mathematical and Statistical Psychology*, 45, 31–49.

Bodington, J. C. (2015a). Evaluating wine-tasting results and randomness with a mixture of rank preference models. *Journal of Wine Economics*, (10)1, 31–46.

Bodington, J. C. (2015b). Testing a mixture of rank preference models on judges' scores in Paris and Princeton. *Journal of Wine Economics*, 10(2), 173–189.

Bodington, J. C. (2017a). Wine, women, men and type II error. *Journal of Wine Economics*, 12(2), 161–172.

Bodington, J. C. (2017b). The distribution of ratings assigned to blind replicates. *Journal of Wine Economics*, forthcoming.

Cao, J., and Stokes, L (2017). Comparison of different ranking methods in wine tasting. *Journal of Wine Economics*, 12(2), 203–210.

Chen, W. (2014). *How to order sushi.* PhD diss., Harvard University.

Cicchetti, D. (2014). *Blind Tasting of South African Wines: A Tale of Two Methodologies.* American Association of Wine Economists, Working Paper No. 164.

Cicchetti, D. (2017). Evaluating the value of triplicate tastings of a given wine: Biostatistical considerations. *Journal of Wine Research*, 28(2), 135–143.

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.
Cortez, P., Cerdeira, A., Almeida, F., Matos, T., and Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, 47(4), 547–553.
Critchlow, D. E. (1980). Metric methods for analyzing partially ranked data. *Lecture notes in Statistics 34*. New York: Springer-Verlag.
Frost, M. B., and Nobel, A. (2002). Preliminary study of the effect of knowledge and sensory expertise on liking for red wines. *American Journal of Enology and Viticulture*, 53(4), 275–284. See also related *UC Davis Viticulture and Enology*, Summary 115.
Ginsburgh, V., and Zang, I. (2012). Shapley ranking of wines. *Journal of Wine Economics*, 7 (2), 169–180.
Greene, W. H., and Hensher, D. A. (2010). *Modeling Ordered Choices*. Cambridge: Cambridge University Press.
Green, P. E., and Rao, V. (1972). *Applied Multidimensional Scaling: A Comparison of Approaches and Algorithms*. Holt, Rienhart and Winston, Austen.
Hastings, J. S., Kane, T. J., and Staiger, D. O. (2006). *Preferences and heterogeneous treatment effects in a public school choice lottery*. National Bureau of Economic Research, Working Paper 12145.
Hodgson, R. T. (2008). An examination of judge reliability at a major U.S. wine competition. *Journal of Wine Economics*, 3(2), 105–113.
Hodgson, R., and Cao, J. (2014). Criteria for accrediting expert wine judges. *Journal of Wine Economics*, 9(1), 62–74.
Keane, M. P., and Wasi, N. (2013). *The structure of consumer taste heterogeneity in revealed vs. stated preference data*. University of Oxford, Nuffield College, Economics Papers from Economics Group, No. 2013-W10.
Mallows, C. L. (1957). Non-null ranking models. *Biometrika*, 44, 114–130.
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 and Hall.
Mc Breen, J., Goffette-Nagot, F., and Jensen, P. (2009). *An agent-based simulation of rental housing markets*. Groupe d'Analyse et de Théorie Economique: Working Paper GATER 2009-08.
Nachev, A., and Hogan, M. (2013). *Using data mining techniques to predict product quality from physiochemical data*. Business Information Systems, Cairnes Business School, NUI, Galway, Ireland.
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.
Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine*, 50(302), 157–175.
Plackett, R. L. (1975). The analysis of permutations. *Applied Statistics*, 24, 193–202.
Rajan, U., and Sinha, A. (2008). *Equilibria in a hotelling model: First-mover advantage?* Ross School of Business, Working Paper No. 1114.
Rhee, D., de Palma, A., and Thisse, J. (1998). First-mover disadvantage with consumers' idiosyncratic preferences along unobservable characteristics. *Regional Science and Economics*, 36(1), 99–117.
Theusen, K. F. (2007). *Analysis of ranked preference data. Informatics and mathematical modeling*. Master's thesis, Technical University of Denmark, Kongens Lyngby, Denmark.

Train, K. (2002). *Discrete Choice Methods with Simulation*. Cambridge: Cambridge University Press.

Vargo, M. D. (1989). *Microbiological spoilage of a moderate acid food system using a dairy-based salad dressing model.* Master's thesis, Department of Food Science and Nutrition, The Ohio State University. See also discussion in Fligner, M. A., and Verducci, J. S. (1993), *Probability Models and Statistical Analyses for Ranking Data*. Springer-Verlag, 11–14.

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.

Wine and Spirit Education Trust. (2014). *Wines and spirits, looking behind the label*. London: Wine and Spirit Education Trust.

Yue, C., Zhao, S., and Kuzma, J. (2015). Heterogeneous consumer preferences for nanotechnology and genetic-modification technology in food products. *Journal of Agricultural Economics*, 66, 308–328.