Journal of Wine Economics, Volume 7, Number 2, 2012, Pages 155–168
doi:10.1017/jwe.2012.24

A Graphical and Statistical Analysis of the *Judgment of Princeton* Wine Tasting*

Daniel L. Ward [a]

Abstract

The wine ratings from the a wine tasting comparing wines from a young wine region, New Jersey, to wines from Bordeaux was held June 8, 2012 in Princeton, NJ. Graphical analysis revealed substantial differences in the use of the rating scale between judges, both in the centroids of their scores and the variability in their scores. Analysis of variance failed to detect any significant differences in the white wines regardless of data transformation or statistical model. Analysis of raw data from the red wines suggested significant differences, but severely violated ANOVA model assumptions and was invalid. Rank transformation, standardization, a model with heterogeneous variances, and Friedman's test all indicated no significant differences among red wines. Based on the confidence interval on the difference between all New Jersey and all Bordeaux wines in each flight it was estimated that differences larger than 1.5 on the 20-point scale would have been declared significantly different for either white or red wines. Therefore, the tasting was powerful enough that any meaningful differences between the wines from the two origins would have been detected. (JEL Classification: C19, Q19)

Keywords: data analysis, wine judges, heterogeneous variances, New Jersey, Friedman's test, Judgment of Princeton.

I. Introduction

The famous 1976 wine tasting in France that thrust California wines onto the world stage served as the model for a tasting conducted June 8, 2012 in Princeton, New Jersey. Just as the 1976 tasting ("Judgment of Paris" (Taber, 2006)) had compared wines from a nascent California industry with those of well-established French winemakers, the 2012 tasting ("Judgment of Princeton" (JOP)) compared wines from another young wine region, New Jersey, with respected wines from Bordeaux. To elucidate and clarify the results from the recent wine tasting we have undertaken

*Thanks to the editor, Karl Storchmann, and an anonymous reviewer who provided helpful editorial advice.

[a] Rutgers University, New Jersey Agricultural Experiment Station, 121 Northville Road, Bridgeton, NJ 08302. Email: Dward@AESOP.Rutgers.edu.

https://doi.org/10.1017/jwe.2012.24 Published online by Cambridge University Press

an analysis of the tasting grades from the JOP with the objectives of testing for differences among the wines, estimating effects of sitting at the same table, and testing for differences between regions. The results indicate that the New Jersey and French wines do not, on average, differ in quality; and that the judges showed no signs of bias due to table seating.

The analysis published on the Liquid Asset website follows the procedure suggested by Quandt (2006) and implemented in his WineTaster® software. The analysis presented also includes considerable examination of the judges for consistency and correlation. Our analysis in this paper focuses on the wines and handles judge effects as a nuisance variable. The testing procedure recommended by Quandt (2006) for comparing the wines is essentially Friedman's test followed by pair-wise comparisons. In their nomenclature, the rank sums are called "points against" and the judges are the blocks. Ranking the scores by judge overcomes some of the problems caused by judges using the scales differently, but at the cost of much of the information in the scores. Ranking provides the order information but much of the information about distance between scores assigned is lost. To preserve this information we present and examine an analysis using a transformation based on standardized scores, as well as a heterogeneous variances model instead of the rank-based Friedman procedure.

Biases induced by the conduct of the tasting could enter into the scoring without the knowledge of the judges or organizers. Any discussion or facial cues from other tasters at the same table could bias a judge's score. The effect of the table at which the judges were seated could create a clustering of scores. Using the seating arrangement, we perform an *ad hoc* analysis to estimate the variance due to table.

It may be inferred that an underlying objective in conducting the JOP was to determine whether the collection of wines from New Jersey were of significantly lower or higher quality as scored by the expert judges. Comparing individual wines and ranking them does not answer this question directly, especially when the precision of a tasting is insufficient to separate all the wines. We have addressed this question within each flight with a single-degree-of-freedom contrast comparing the origins, which is, comparing the mean of all the New Jersey wines to the mean of all the Bordeaux wines.

II. Method

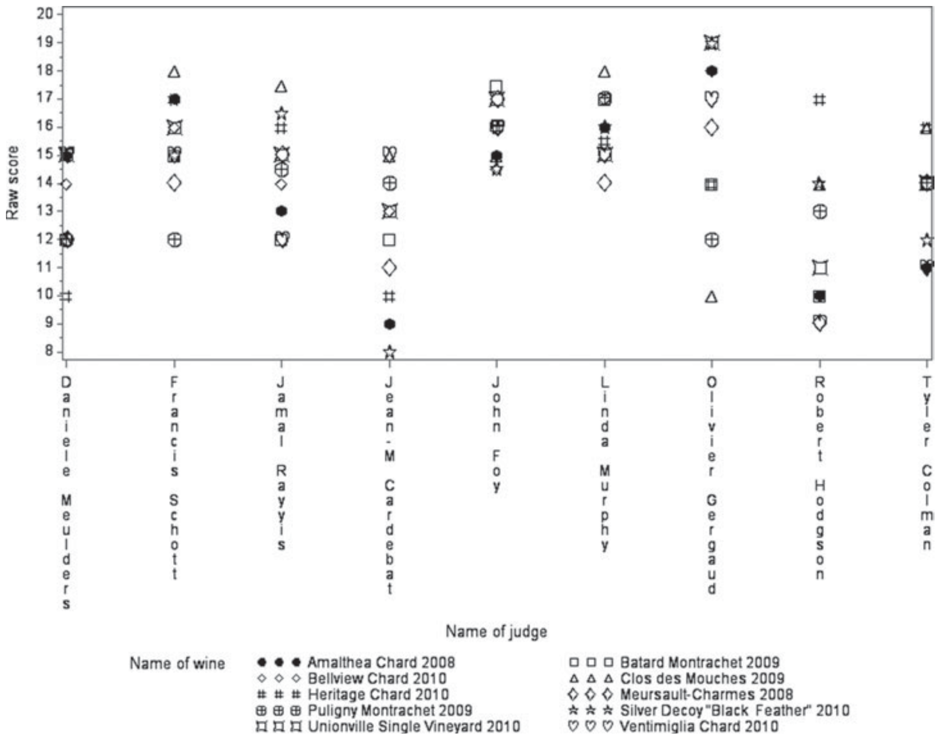
The data was retrieved from the Liquid Assets web page reporting the tasting (http://www.liquidasset.com/report161.html, last accessed 11 August 2012).[1] In our

[1] The raw results are also reported in the Editorial to this issue of the Journal of Wine Economics (Ashenfelter and Storchmann, 2012).

https://doi.org/10.1017/jwe.2012.24 Published online by Cambridge University Press

Figure 1

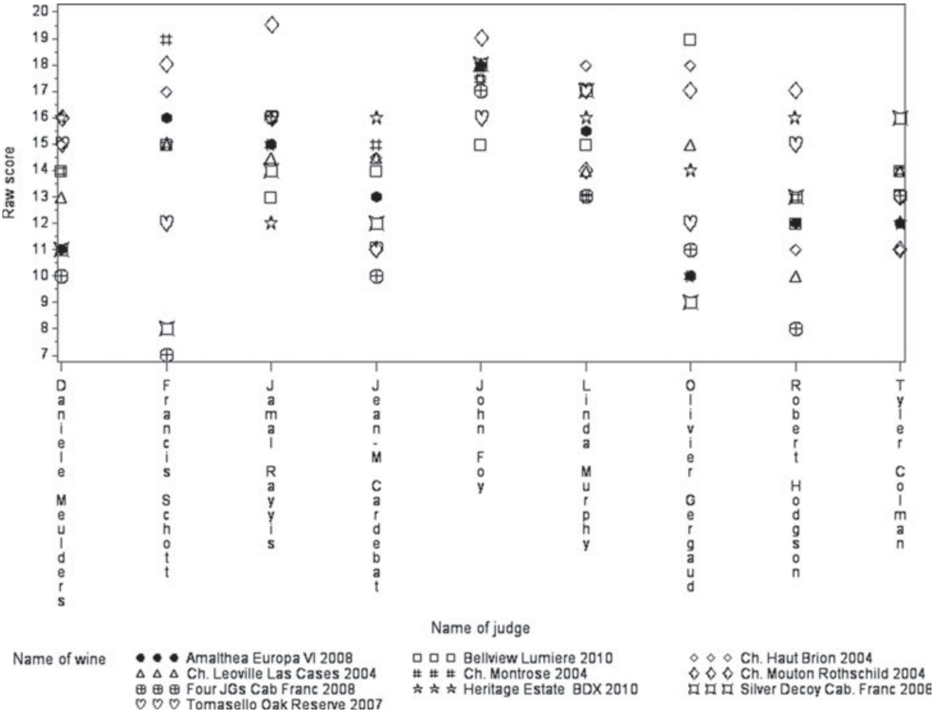
Tasting Scores of the 10 White Wines in the Judgment of Princeton Wine Tasting by Wine and Judge



presentation the revealed judges' and wines' names are preserved (judges' initials are used in text) as they were presented on the Liquid Assets website. The judges' scores on the 0–20 point scale were used as the raw data for all the analyses. The graphical analysis presented of the raw data, mean-centered data, ranked data (ranked within judge), and standardized data motivates the statistical analyses. A general linear model approach is taken starting from an assumed two-factor mixed model (with judge and wine as the factors), to two-factor mixed models on transformed data, and ending with a two-factor mixed model with heterogeneous variances for judges. Throughout judge is considered a random effect. The Mixed procedure (Littell et al., 2006) of the SAS System was used for the calculations. The restricted maximum likelihood estimation performed by the Mixed procedure was used for all models and the type three tests of fixed effects (analogous to the partial sums of squares) were interpreted. Pair-wise comparisons of all wines within a flight were performed using the unprotected least significant difference (LSD) test. Friedman's test, as recommended by Quandt (2006), was also performed (also with uncorrected

Figure 2

Tasting Scores of the 10 Red Wines by Wine and Judge

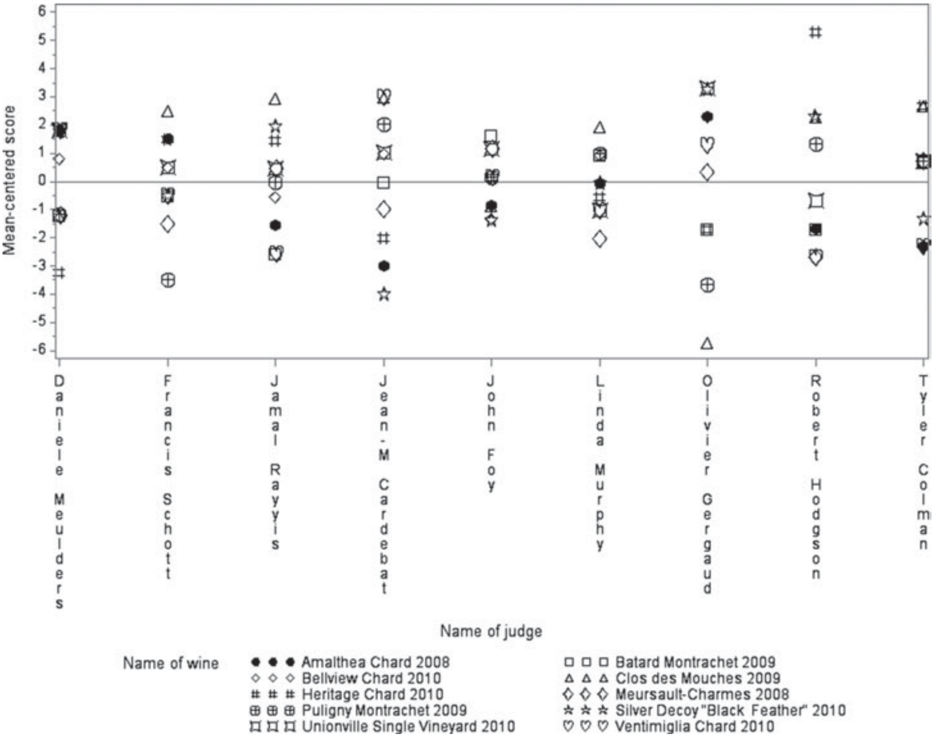


pair-wise comparisons) to permit comparison of the other methods with a rank randomization alternative.

Separate exploratory statistical analyses were performed to: 1) estimate variance components including the effect of table (table at which the judge was seated); 2) compare the mean of New Jersey wines to the mean for the Bordeaux wines. Variance components were estimated using restricted maximum likelihood estimation with all factors considered nested random effects (wine within judge and judge within tables) for the white and red flights separately. The intraclass correlation coefficients were then calculated as the ratio of individual variance components to the sum of all the variances. A comparison of the means of all the wines from each origin was conducted for each flight separately using a single-degree-of-freedom contrast within the context of the different linear models used for comparing individual wines. Confidence intervals on the differences calculated by these contrasts are presented. Because the models have different power, the confidence intervals provide a means of evaluating the ability of the model to detect meaningful differences in the scores.

Figure 3

Mean-Centered Tasting Scores of the 10 White Wines by Wine and Judge

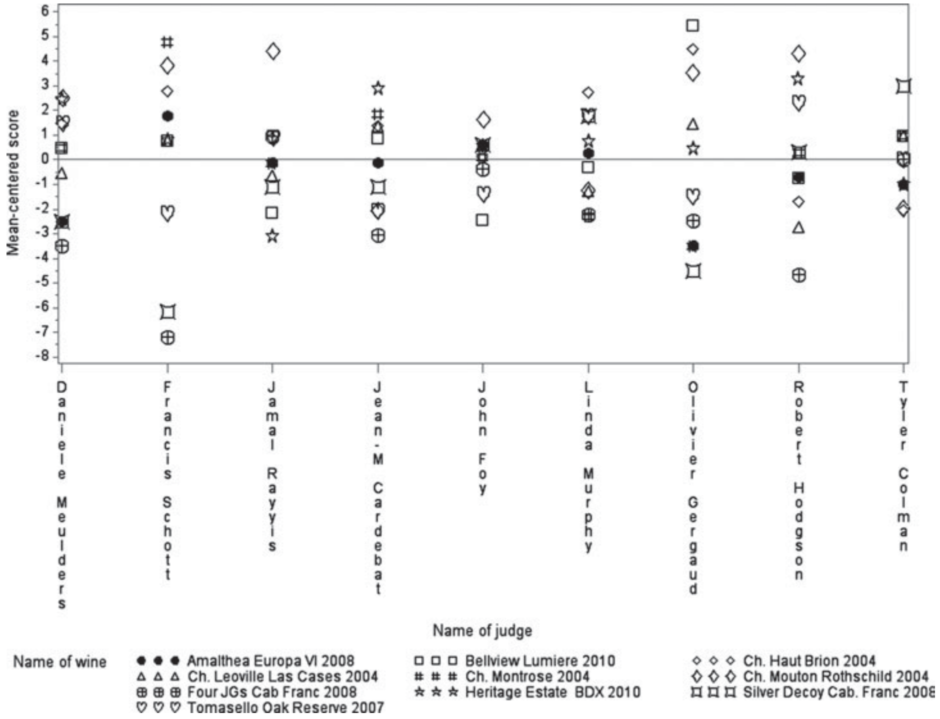


III. Results and Discussion

To appreciate the variability among judges in both how high they scored the wines and the way in which they used the 20-point scale we inspect Figure 1 for the white wines and Figure 2 for the red wines. From these Figures, it is apparent that some judges scored the wines much higher overall than other judges did. Judges that use different regions of the 20-point scale can have more impact on overall means for each wine, giving them unequal influence on the results. Mean-centering (subtracting the judge's mean score from all their scores) transformation of the scores neatly corrects for this inequality (Figure 3 and Figure 4). After mean centering, the variation in the judges' scores can be readily assessed visually. A two-factor linear model including judge as an effect would essentially be partitioning the variability displayed in these figures into wine effect and residual. A simple two-factor model though would assume homogeneity of variances among judges. Differences in variances of judges' scores appear to invalidate models assuming equal variances.

Figure 4

Mean-Centered Tasting Scores of the 10 Red Wines by Wine and Judge



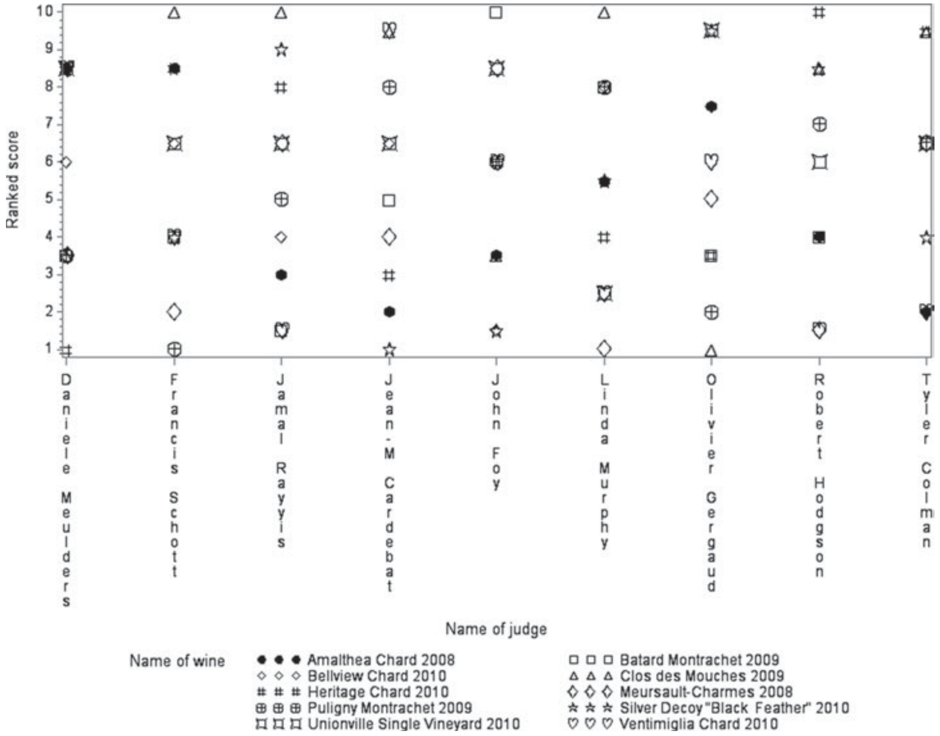
Rank transformation, or similarly Friedman's test, corrects the inequality in mean scores neatly, but at the same time it discards the information about variability in judges' scores (Figure 5 and Figure 6). The rank transformation also stabilizes the variance within judges, but fine distinctions in the tasters scoring are lost. For example, F.S. scores the two Cabernet Franc varietal wines substantially lower than most of the other wines (Figure 2). This distinction is lost after ranking (Figure 6).

Also important is that the range of scores is quite different for the different judges. For example, judge J.F. scores all the white wines between 14.5 and 17.5, whereas J.C. scored them from 8 to 15, a greater than two-fold difference in range. Whether the difference in range is due to differences in discriminatory ability, differences in how the judges use the 20-point scale, or differences in taste preferences of the judges is beyond our knowledge, but the fact that some judges separate the wines more is important to our goal. Standardization by dividing each score by the standard deviation of all the scores from the same judge was conducted in an attempt to ensure that each judge contributed equally to the

https://doi.org/10.1017/jwe.2012.24 Published online by Cambridge University Press

Figure 5

Ranked Scores of the 10 White Wines by Wine and Judge



overall score while preserving more of the information in their scores than ranking would. After standardization, the scores have much more uniformity in variances across judges while retaining more than just order information (Figure 7 and Figure 8).

The effect of which table the judge was seated was explored with an *ad hoc* analysis to allay any fears of bias from this effect. For each flight of wines (whites and reds), variance components were estimated assuming that all effects were random effects and mutually uncorrelated and that judges were nested within tables. The variance component model used was:

$$\text{score}_{\text{ijr}} = \mu + \alpha_{\text{i}} + \beta_{\text{ij}} + \varepsilon_{\text{ijr}}$$

where:

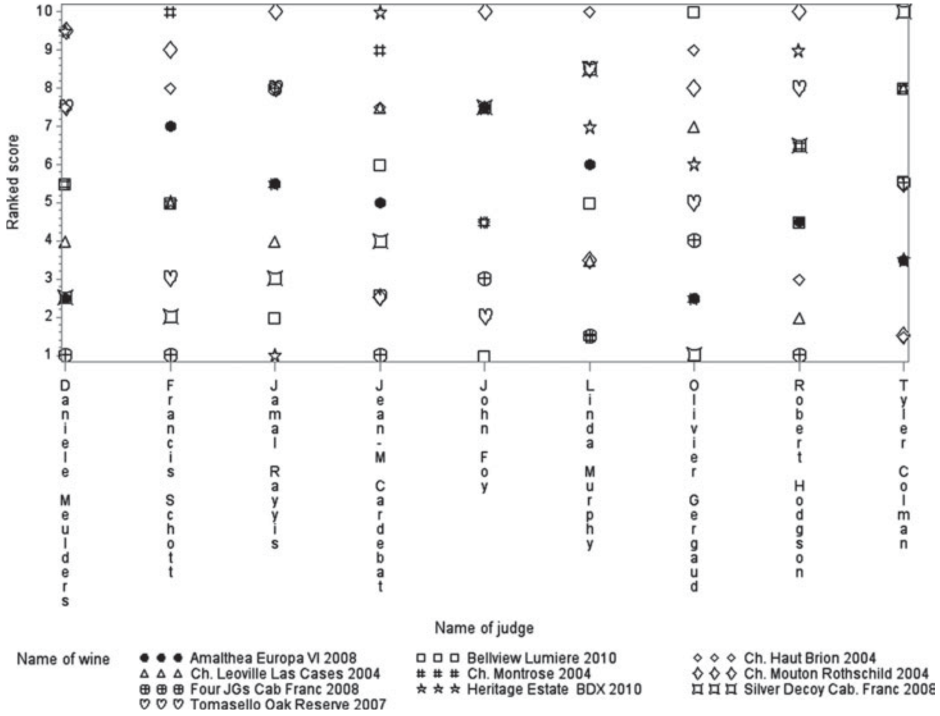
μ is the overall mean of the population,

α_{i} is the random effect of table,

https://doi.org/10.1017/jwe.2012.24 Published online by Cambridge University Press

Figure 6

Ranked Scores of the 10 Red Wines by Wine and Judge



β_{ij} is the random effect of judge within a level of table,

ε_{ijr} is the random effect of wine within a level of judge, within a level of table.

The variance components for the white and red wines suggest that little of the variation in scores is explained by the table-to-table variation (Table 1). For the white wines judge-to-judge variation accounted for 37% of the total variance, while for red wines judge-to-judge variation only accounted for 12% of the total. For both reds and whites, wine-to-wine variation explained the majority of the total variation. Being seated at the same table had no appreciable effect on scoring.

For the white wines the omnibus test of equality of means indicates that there was no significant effect of wine regardless of the transformation or statistical model used (Table 2). The unadjusted comparisons are overly liberal at detecting differences in a tasting of this size. The pair-wise comparisons are unadjusted for multiplicity so making 45 comparisons at $\alpha = 0.05$ we would expect to find two differences significant due to chance alone. None of

Table 1

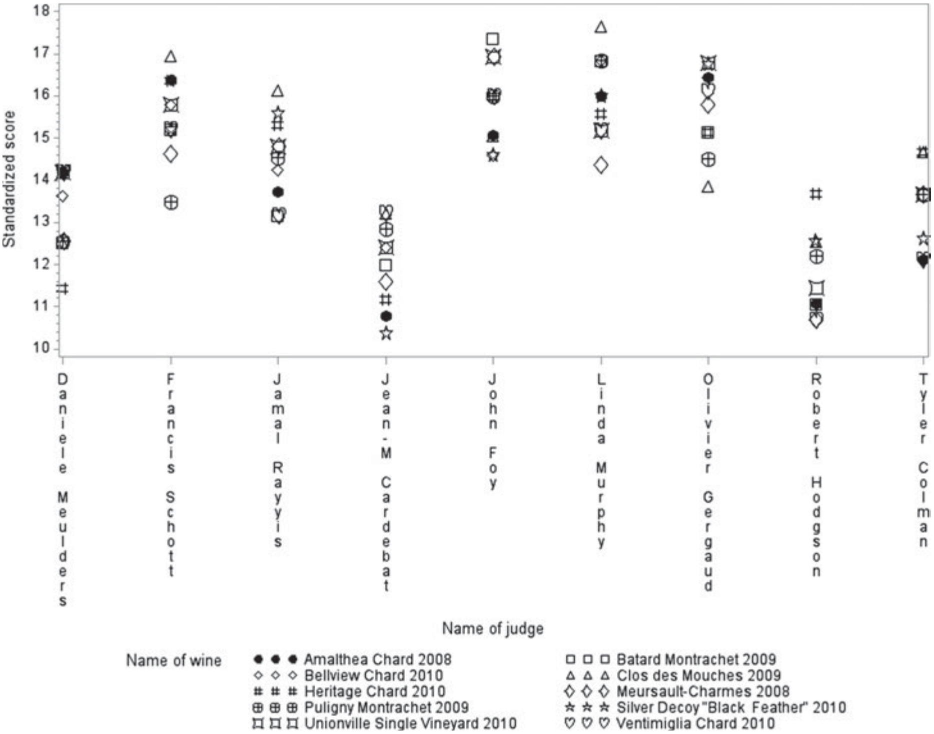
Variance Components for Wine Tasting Scores of White and Red Wines

	Variance component				
Wine class	*Table*		*Judge*		*Wine (residual)*
White wines	0.0	(0.0)	2.5	(0.571)	4.3
Red wines	0.4	(0.144)	0.9	(0.150)	6.3

Table, Judge and Wine (residual) were all considered random effects for this nested random effects model. The intraclass correlation coefficients are in parentheses.

Figure 7

Standardized Scores of the 10 White Wines by Wine and Judge



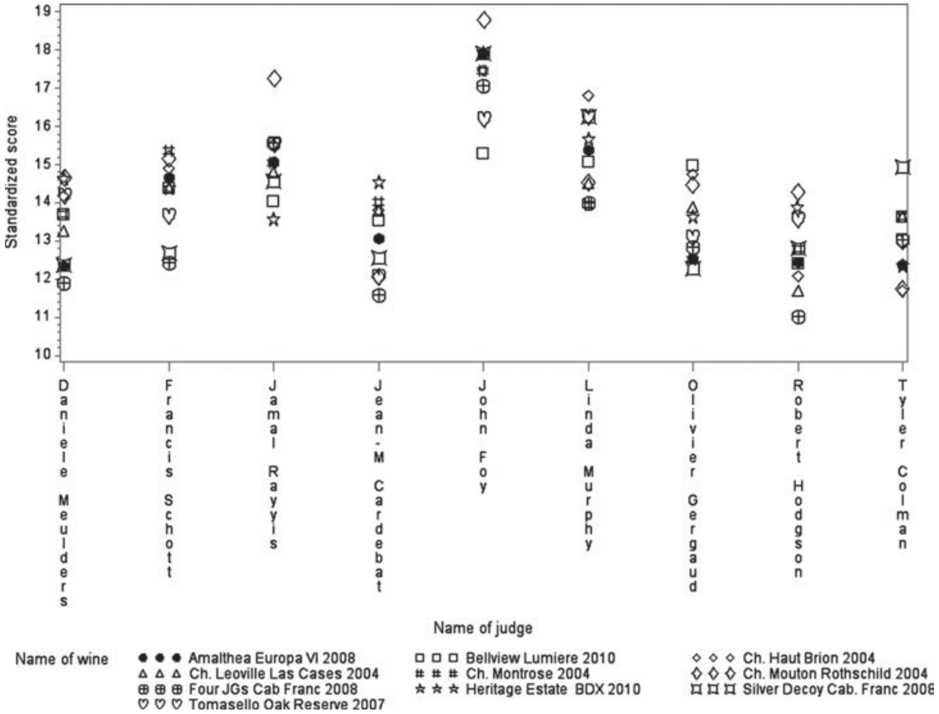
Raw scores were divided by the standard deviation of all the scores from the same judge.

the linear models detected more than two significant pair-wise differences, which is consistent with the omnibus test indication of no significant effect of wine. An overall conclusion emerges from the results of all the data transformation and modeling techniques: that the average scores for the white wines are not different.

https://doi.org/10.1017/jwe.2012.24 Published online by Cambridge University Press

Figure 8

Standardized Scores of the 10 Red Wines by Wine and Judge



Raw scores were divided by the standard deviation of all the scores from the same judge.

Analysis of the raw or centered scores for the red wines indicated that there were significant differences (Table 3, omnibus test). From this, the unwary analyst could be induced to interpret the pair-wise comparisons such that the top rated wine was significantly better than the bottom three wines, the wine rated second was better than the bottom two wines and the top seven wines were all better than the wine rated lowest. However, the assumptions of these tests are severely violated by the data rendering these results invalid. Given the large number of pair-wise comparisons in this tasting, without correcting for multiplicity the analyses using the rank transformed data, heterogeneous variances model, or Friedman's test would all drastically inflate the type I error in interpreting the pair-wise comparisons. The omnibus test on the ranked data, standardized data, heterogeneous variances model, and Friedman's procedure all indicate that there were no significant differences among the wines. An overall conclusion again emerges from the results of the ranked data transformation, heterogeneous variances model and Friedman's test: the average scores for the red wines are not different.

We have also reduced the JOP to just two questions: 1) Is the quality of the white wines from New Jersey significantly different from that of the wines

Table 2

Summary of Scores Assigned to the 10 White Wines by Wine and Transformation or Statistical Method

Wine	*Raw*		*Ranked*		*Centered*		*Standardized*		*Heterogeneous*		*Friedman's*	
Clos des Mouches 2009	15.1	a	3.7	a	0.850	a	14.7	a	15.8	a	33.5	a
Unionville Single Vineyard 2010	15.0	a	4.2	ab	0.794	a	14.6	a	14.6	ab	38.0	ab
Heritage Chardonnay 2010	14.6	a	5.1	ab	0.406	a	14.4	a	14.8	ab	45.5	abc
Silver Decoy "Black Feather" 2010	14.4	a	5.3	ab	0.239	a	14.2	a	14.2	ab	47.5	abc
Bellview Chardonnay 2010	14.2	a	5.9	ab	– 0.039	a	14.1	a	14.0	ab	53.0	bc
Ventimiglia Chardonnay 2010	13.9	a	6.4	ab	– 0.317	a	14.0	a	13.4	b	57.5	c
Puligny Montrachet 2009	13.8	a	5.8	ab	– 0.372	a	14.1	a	14.1	ab	52.0	bc
Amalthea Chardonnay 2008	13.8	a	6.1	ab	– 0.428	a	14.0	a	13.7	ab	54.5	bc
Batard Montrachet 2009	13.7	a	5.9	ab	– 0.483	a	14.1	a	14.2	ab	53.0	bc
Meursault-Charmes 2008	13.6	a	6.7	b	– 0.650	a	13.9	a	13.6	b	60.5	c
Omnibus P-value	0.8027		0.5251		0.7476		0.7323		0.5340		0.5476	

The transformations are; Raw = raw data, Ranked = data ranked within judges, Centered = mean of judge's scores subtracted from each score, and Standardized = scores divided by standard deviation of all scores from that judge. Heterogeneous denotes that a two factor mixed effects model with separate intercepts and variances for each judge was used, while the column denoted Friedman's indicates that Friedman's test was used. Estimates with no letters in common are significantly different at $\alpha = 0.05$ according to multiple comparisons of all pair-wise differences conducted using the Least Significant Difference test (LSD, without experiment-wise error protection). The omnibus null hypothesis is that the mean scores (or for Friedman's the rank sums) for all wines are equal.

https://doi.org/10.1017/jwe.2012.24 Published online by Cambridge University Press

Table 3

Summary of Scores Assigned to the 10 Red Wines by Wine and Transformation or Statistical Method

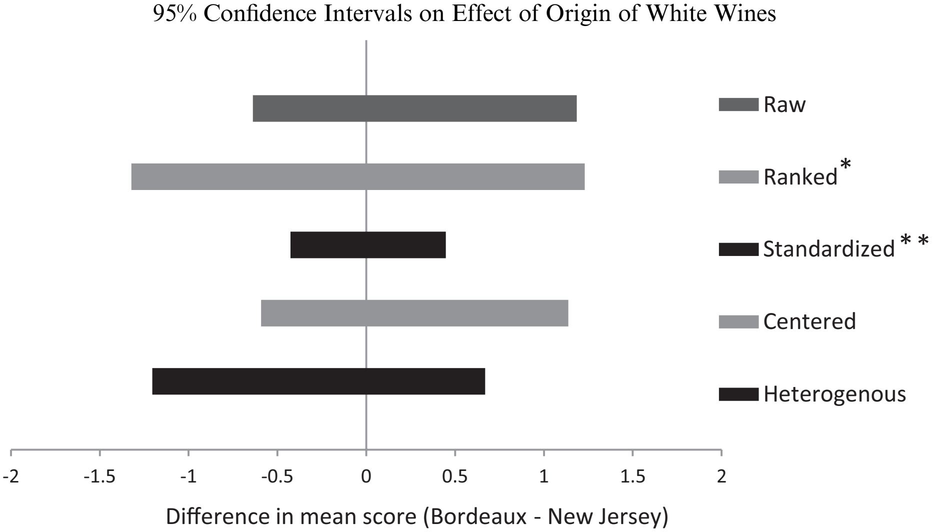
Wine	*Raw*		*Ranked*		*Centered*		*Standardized*		*Heterogeneous*		*Friedman's*	
Ch. Mouton Rothschild 2004	15.8	a	3.9	a	1.639	a	14.8	a	15.5	a	35.0	a
Ch. Haut Brion 2004	15.3	ab	4.4	a	1.139	ab	14.6	a	15.2	a	40.0	ab
Heritage Estate BDX 2010	15.0	abc	4.5	a	0.806	abc	14.5	a	15.3	a	40.5	ab
Bellview Lumiere 2010	14.6	abc	5.8	ab	0.361	abc	14.1	ab	14.0	ab	52.0	ab
Ch. Montrose 2004	14.5	abc	5.1	a	0.306	abc	14.3	a	14.4	ab	46.0	ab
Ch. Leoville Las Cases 2004	14.2	abc	5.6	ab	0.028	abc	14.2	a	14.2	ab	50.5	ab
Tomasello Oak Reserve 2007	14.1	abc	5.4	ab	– 0.083	abc	14.2	a	14.4	ab	49.0	ab
Amalthea Europa VI 2008	13.6	bcd	6.1	ab	– 0.583	bcd	14.0	ab	13.7	ab	55.0	bc
Silver Decoy Cab. Franc 2008	13.1	cd	6.0	ab	– 1.083	cd	14.0	ab	13.7	ab	54.0	bc
Four JGs Cab Franc 2008	11.7	d	8.1	b	– 2.528	d	13.2	b	12.1	b	73.0	c
Omnibus P-value	0.0232		0.1797		0.0122		0.1061		0.2037		0.2039	

The transformations are; Raw = raw data, Ranked = data ranked within judges, Centered = mean of judge's scores subtracted from each score, and Standardized = scores divided by standard deviation of all scores from that judge. Heterogeneous denotes that a two factor mixed effects model with separate intercepts and variances for each judge was used, while the column denoted Friedman's indicates that Friedman's test was used. Estimates with no letters in common are significantly different at $\alpha = 0.05$ according to multiple comparisons of all pair-wise differences conducted using the Least Significant Difference test (LSD, without experiment-wise error protection). The omnibus null hypothesis is that the mean scores (or for Friedman's the rank sums) for all wines are equal.

https://doi.org/10.1017/jwe.2012.24 Published online by Cambridge University Press

Figure 9

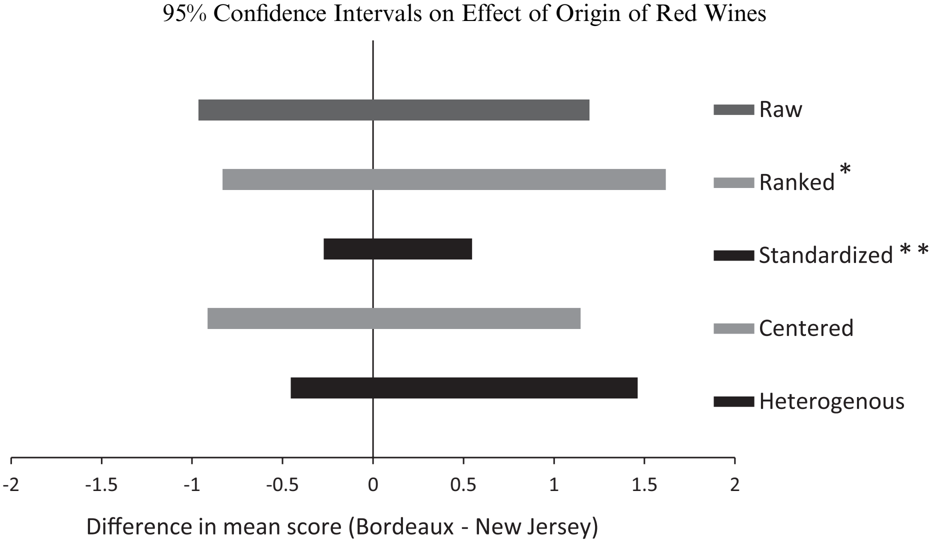
Confidence Intervals on the Mean Difference of Wine Tasting Scores of White Wines from Bordeaux or New Jersey



*Ranked data are in units of ranks while units for Raw, Centered, and Heterogeneous are points on a 1–20 scale. **Standardized data are in units of standard deviations of the individual judge's scores on a 1–20 point scale.

Figure 10

Confidence Intervals on the Mean Difference of Wine Tasting Scores of Red Wines from Bordeaux or New Jersey



from Bordeaux; and 2) The same question for the red wines. A single-degree-of-freedom contrast was used to address these questions and was non-significant ($\alpha = 0.05$) for both red and white wines. The confidence intervals on the difference between the mean scores for the New Jersey and Bordeaux wines provide a measure of how large the difference would have needed to be in order to be significant based on this tasting (Figure 9 and Figure 10). A difference in the means of the raw scores larger than 1.5 would have been declared significantly different for either white or red wines. It seems unlikely that such a small difference could be considered meaningful and suggests that the tasting was powerful enough that any meaningful differences between the wines from the two origins would have been detected.

The JOP provided a venue for comparing a selected sample of wines from an established wine region known for its excellent quality with wines from a developing wine region aspiring to excellence in wine quality. The tasting revealed that, after accounting for differences among judges, the differences between the wines or the wine regions were too small to detect. These select New Jersey wines were found to be not dissimilar in quality from the excellent Bordeaux wines as ranked at this tasting.

References

Ashenfelter, O. and Storchmann, K. (2012). Editorial: the *Judgment of Princeton* and other papers. *Journal of Wine Economics*, 7(2), 139–142.

Littell, R. C., Milliken, G.A., Stroup, W.W., Wolfinger, R.D. and Schabenberger, O. (2006). *SAS® for Mixed Models*, 2nd ed. Cary, NC: SAS Institute Inc.

Quandt, R.E. (2012). Tasting Report No. 161. *Liquid Assets*. Online at http://www.liquidasset.com/report161.html (last accessed August 22, 2012).

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

Taber, G. (2006). *Judgment of Paris: California vs. France and the Historic 1976 Paris Tasting That Revolutionized Wine*. New York, NY: Scribner.

https://doi.org/10.1017/jwe.2012.24 Published online by Cambridge University Press