Journal of Wine Economics, Volume 3, Issue 2, Fall 2008, Pages 105–113

An Examination of Judge Reliability at a major U.S. Wine Competition*

Robert T. Hodgson[a]

Abstract

Wine judge performance at a major wine competition has been analyzed from 2005 to 2008 using replicate samples. Each panel of four expert judges received a flight of 30 wines imbedded with triplicate samples poured from the same bottle. Between 65 and 70 judges were tested each year. About 10 percent of the judges were able to replicate their score within a single medal group. Another 10 percent, on occasion, scored the same wine Bronze to Gold. Judges tend to be more consistent in what they don't like than what they do. An analysis of variance covering every panel over the study period indicates only about half of the panels presented awards based solely on wine quality. (JEL Classification: Q13, Q19)

I. Introduction

In the spring of 2003 the author contacted the chief judge of the California State Fair wine competition in Sacramento, proposing an independent analysis of the reliability of its judges. The following questions were asked. Why is it that a particular wine wins a Gold medal at one competition and fails to win any award at another? Is this caused by bottle-to-bottle variability of the wine? To what extent is the variability caused by differing opinions within a panel of judges? Finally, could the variability be caused by inability of individual judges to reproduce their scores? As pointed out by Margaret Cliff and Marjorie King (1997) "Wine judges are rarely, if ever, subjected to rigorous testing. Neither are competitions based on rigorous experimental design with replication to examine judge reproducibility."

* I would like to thank the administration and advisory board of the California State Fair Wine Competition for supporting this research and agreeing to release the results. Taking such a leadership role benefits the entire wine industry. I would especially like to thank G.M. "Pooch" Pucilowski, chief judge, and Kem Pence, wine department chairperson of the California State Fair Commercial Wine Competition, for their continued support of this study. In addition, Matt Sainson, www.ijudgewine.com, was the programmer responsible for data management for the entire competition. I am also indebted to an anonymous referee.

[a] Professor Emeritus, Depar tment of Oceanography, Humboldt State University, Arcata, CA 95521, email: bob@fieldbrookwinery.com

The California State Fair hosts the oldest commercial wine competition in North America. Entries are limited to wines produced in California and its judges are selected from a broad range of wine professionals: winemakers, wine buyers, wine critics and professors of enology and viticulture. Lima (2006) calculates a positive relationship between price and medal status at the State Fair wine competition so awards at the State Fair can be important to a winery's economic health.

The lack of concordance[1] between judges is well known (Ashenfelter, 2006), but an investigation of individual judge consistency has not been published. Ashenfelter and Quandt (1999) and Cicchetti (2004a) appear to disagree on the benefits of rank order statistics (a non parametric vs. parametric basis) to evaluate wine tastings. However, both approach the reliability of judges based on the principle of concordance, i.e., good judges agree with each other, whether by score or by rank. Basing reliability on concordance rather than consistency may be attributed to psychological research, where in order to obtain independent observations, ratings of identical subjects are not possible. Thus, the seminal papers by Bartko (1966) and Shrout and Fleiss (1979) on Intraclass Correlations, which is the approach taken by Cicchetti (2004a, 2006) on evaluating the famous 1976 Paris tasting, base reliability on concordance. This paper advances the thesis that consistency is more fundamental than concordance in evaluating judge reliability because it is a better basis of measuring experimental error.

II. Methods

The results that follow are based on four triplicate samples served to 16 panels of judges. A typical flight consists of 30 wines. When possible, triplicate samples of all four wines were served in the second flight of the day randomly interspersed among the 30 wines. A typical day's work involves four to six flights, about 150 wines. Each triplicate was poured from the same bottle and served on the same flight. The overriding principle was to design the experiment to maximize the probability in favor of the judges' ability to replicate their scores.

The judges first mark the wine's score independently, and their scores are recorded by the panel's secretary. Afterward the judges discuss the wine. Based on the discussion, some judges modify their initial score; others do not. For this study, only the first, independent score is used to analyze an individual judge's consistency in scoring wines.

III. Data

Approximately 3,000 wines per year were entered during the evaluation period. A typical data set for one panel (2006) is produced below as Table 1. J1 to J4 represent the four judges. The three values associated with R1 represent the scores given the first group of

[1] In this study, *concordance* is used to describe agreement between judges; *consistency* will be used to describe the ability of individual judges to repeat their scores on identical wines. *Reliability* is loosely defined to be a combination of both but is often equated with concordance in the cited literature.

Table 1
Example of Recorded Data for Panel Q (2006)

Judge Code	*Original Score*			
	J1	J2	J3	J4
R1	84	90	80	80
R1	84	88	94	82
R1	80	80	82	86
R2	80	80	84	84
R2	90	90	80	82
R2	96	80	80	82
R3	80	80	80	80
R3	80	80	80	80
R3	80	80	80	80
R4	88	96	80	80
R4	90	96	82	88
R4	96	90	80	84

Individual statistics are calculated for each judge including four values of range and standard deviation. Also calculated for each judge are their maximum range (maximum error in consistency) and a pooled standard deviation (a typical error in consistency). A two-way Analysis of Variance (ANOVA) was performed for each panel to determine the relative importance of a wine's score between a "wine factor" and the "judge factor." The calculations are performed using Microsoft Excel.

replicate wines, the three for R2, the second, etc. Judges are only asked to provide letter scores, Bronze+, Silver–, etc. Letter scores are later converted to numerical scores ranging from 80 points (No Award) to 100 points.

IV. Results

The results are divided into two categories, individual judge performance and group (panel) performance.

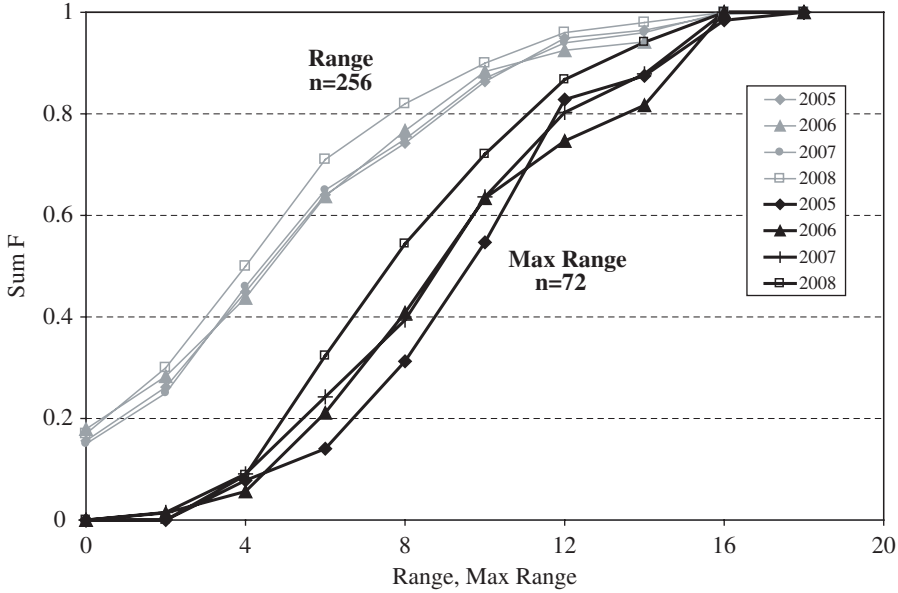
A. *Individual Judge Performance*

Range and Maximum Range

Figure 1 shows cumulative frequency distributions for range and maximum range for the period 2005 to 2008. The median range is about 4 points, e.g., Silver – to Silver +. More interesting is the maximum range where only 10 percent of the judges were *consistently consistent* to a single medal range (a maximum 4-point spread). On the other extreme, another 10 percent awarded scores ranging from Bronze to Gold (a12-point spread) or worse (No Award to Gold). Although the median maximum range varies from year to year, it is typically about 8 points corresponding to, e.g., Bronze+ to Gold–.

Figure 1

Cumulative Frequency Distribution for Range and Maximum Range 2005–2008



Examining figure 1, judges were perfectly consistent (0-point spread) about 18 percent of the time. However, this usually occurred for wines that were rejected. That is, when the judges were very consistent, it was often for wines that they did not like.

Pooled standard deviation

A pooled standard deviation is calculated for each judge based on the four triplicate samples. Since there are two degrees of freedom associated with each of the four replicates the pooled standard deviation is

$$s_p=[\{s^2(R_1) + s^2(R_2) + s^2(R_3) + s^2(R_4)\}/4]^{1/2}$$

The median and mean pooled standard deviation for all judges from 2005 to 2008 is 3.6 points. The corresponding 95 percent confidence interval is 14 points, which includes almost the entire range of medals possible, e.g., No Award to Gold–, or Bronze to Gold+.

Year-to year-correlation

Do the most consistent judges repeat their performance year to year? Cicchetti (2004a) asks, "(W)ould the most reliable tasters established in a given competition continue as such, or are these initial findings just another chance phenomenon?" He also asks whether the "'best' judges will continue to remain so in future tastings. To the extent that this is true, it opens up the real possibility of using the more consistent judges to train the less consistent ones to become more highly reliable in their evaluations."

To answer these questions, there is sufficient data to perform a correlation analysis for those judges who participated in at least two competitions. For the 2005-2006 competitions, there were 26 common judges. A scatter diagram, shown in Figure 2, would indicate that a judge's superior performance (in consistency) one year does not correlate with superior performance the next.

B. *Panel Variance*

For each panel, an ANOVA was performed to examine the variation of judge's scores. We presume there are two factors that determine the score given a wine: (1) the quality of the wine and (2) the bias of the judge. Does a wine's score depend on the wine, the judge evaluating the wine, both or neither? In addition to judge bias, judge inconsistency will reduce the significance of the wine factor by increasing experimental error.

Table 2 represents the data from one panel in 2008. Table 3 shows the accompanying ANOVA. Since the P-value for the wine is small (less than 0.05) we assume wine quality is a significant factor in the score it received. As the P-value for the judges is greater than 0.05, it means judge bias is not a significant factor in the wine's final score (case 2 above).

Figure 2
Scatter diagram of pooled standard deviation for the 26 judges who participated in 2005 and 2006
Correlation = –0.01

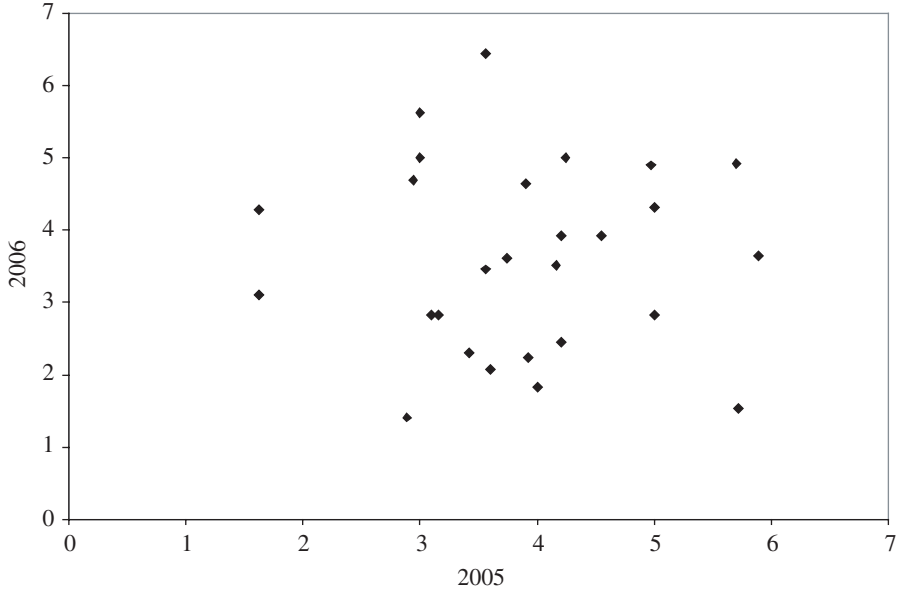


Table 2
Raw Data from Panel xyz, 2008

	J1	J2	J3	J4
R1	84	80	90	96
R1	86	96	96	94
R1	90	84	90	96
R2	80	80	86	84
R2	88	96	92	80
R2	84	80	84	80
R3	84	96	84	80
R3	86	84	84	84
R3	86	84	84	84
R4	86	90	90	88
R4	88	84	94	86
R4	84	84	90	84

Table 3
Analysis of Variance for Data in Table 2

ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Wines	241.7	3.0	80.6	4.4	0.011	2.9
Judges	65.7	3.0	21.9	1.2	0.332	2.9
Interaction	265.7	9.0	29.5	1.6	0.159	2.2
Within	592	32	18.5			
Total	1165	47				

To recognize a good panel, it is desirable to see a small P-value for the wine and a large P-value for the judges, i.e., the wine is important and the judge bias is not. In addition, judge inconsistency increases experimental error, thus reducing the ability of the wine factor to dominate, i.e., it will increase the P-value for the wine. Table 4 lists the P-values for the wines and judges for all panels in the 2008 competition.

Based on the above guidelines, Panels 4, 6, 10, 11, 13 and 15 meet the requirement that it was the wine, not the judge bias, that contributed to the rankings. In Panels 2, 5, 7 and 9, wine quality evidently was a factor although other factors were also important. In panels 1, 3, 8, 12, 14, 16 and 17, the ranking of the wine did not depend so much on the wine as other factors.

Table 4
Summary of ANOVA Results for all Panels, 2008

Panel	*Wine*	*Judges*	*unknown*	*exp error*
1	0.461	0.121	0.000	3.2
2	0.005	0.001	0.046	3.5
3	0.108	0.125	0.210	3.7
4	0.000	0.060	0.208	3.3
5	0.030	0.000	0.288	3.3
6	0.000	0.608	0.055	3.0
7	0.037	0.026	0.041	3.6
8	0.164	0.012	0.017	3.5
9	0.006	0.024	0.000	2.6
10	0.011	0.332	0.159	4.3
11	0.000	0.094	0.060	2.2
12	0.296	0.409	0.012	3.8
13	0.000	0.398	0.020	3.3
14	0.199	0.121	0.204	4.1
15	0.007	0.755	0.149	3.9
16	0.286	0.585	0.281	5.8
17	0.526	0.001	0.477	3.5

Considering all 65 panels tested during 2005 to 2008, Table 5 summarizes their performance into four groups. The top right panel corresponds to cases where the wine factor and not the judge factor was the primary element responsible for the wine's score. The top left panel shows cases where both the wine factor and the judge factor were significant in determining the wine's score. The bottom left panel shows cases where the judge factor was the primary element in determining the wine's score, and the bottom right panel indicates those cases where the wine's score was determined neither by the wine's quality nor judge bias.

In 30 cases, about 50 percent, the wine and only the wine was the significant factor in determining the judges' score. For the remaining 50 percent of the panels, other factors played a significant role in the award received.

Table 5
Summary of ANOVA Analyses 2005–2008

		Judge Factor	*Judge Factor*
		P <.05	P ≥.05
Wine Factor	P <.05	15	30
Wine Factor	P ≥.05	9	11

V. Discussion

The author's interest over the past four years has been to explain the variability in the results of wine competitions. While this data is from a single competition, there is, in the author's mind, no reason to suspect the results are not general. This is because the format of many competitions is similar and because many of these judges participate in other competitions as well. By serving replicate samples on the same flight and poured from the same bottle, the experiment favored optimum performance by the judges. Had the wines been served on different flights, it is reasonable to assume consistency would be less. While this does not directly explain why wines win Golds in some competitions and fail to place in others, it is reasonable to predict that any wine earning *any* medal could in another competition earn *any other* medal, or none at all. Indeed, in 2003 as reported by The Grapevine[2], which tracked over 4,000 wines entering 14 major U.S. wine competitions, more than 1,000 wines receiving a Gold medal in one or more competitions failed to place in others.

To lift their brand above the competition, wineries spent more than $1 million in entry fees at just four California competitions alone this year. The benefit of this expense is the belief by wineries that entry fees offer a valid return on investment: gold medals sell wine. However, a recent article in Wine Business Monthly (Thach, 2008) conducted as a joint effort by 10 global universities with specialties in wine business and marketing found that consumers are not particularly motivated by medals when purchasing wine in retail stores. If consumer confidence is to be improved, managers of wine competitions would be well advised to validate their recommendations with quantitative standards.

Ideally, an examination of the 65 judging panels over four years in Table 5 would show all 65 in the upper right quadrant where wine and only wine is the determinant factor. How can one explain that just 30 panels, less than half, achieved those results? The answers are judge inconsistency, lack of concordance – or both.

VI. Conclusion

The purpose of this investigation was to provide a measure of a wine judge's ability to consistently evaluate replicate samples of an identical wine. With such a measure in hand, it should be possible to evaluate the quality of future wine competitions using consistency as well as concordance with the goal to continually improve reliability and to track improvements associated with procedural changes.

[2] California Grapevine, P.O. Box 22152, San Diego CA 92192 www.calgrapevine.com

References

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

Ashenfelter, O. (2006). Tales from the crypt: Bruce Kaiser tells us about the trials and tribulations of a wine judge. *Journal of Wine Economics*, 1(2), 173–175.

Bartko, J.J. (1966). The intraclass correlation coefficient as a measure of reliability. *Psychological Reports*, 19, 3–11.

Cicchetti, D.V. (2004a). Who won the 1976 wine tasting of French Bordeaux and U.S. cabernets? Parametrics to the rescue. *Journal of Wine Research*, 15, 211–220.

Cicchetti, D.V. (2004b). On designing experiments and analyzing data to assess the reliability and accuracy of blind wine tastings. *Journal of Wine Research*, 15, 221–226.

Cicchetti, D.V. (2006). The Paris 1976 tastings revisited once more: Comparing ratings of consistent and inconsistent tasters. *Journal of Wine Economics*, 1(2), 125–140.

Cliff, M.A. and King, M.C. (1996). A proposed approach for evaluating expert judge performance using descriptive statistics. *Journal of Wine Research*, 7, 83–90.

Cliff, M.A. and King, M.C. (1997). The evaluation of judges at wine competitions: The application of eggshell plots. *Journal of Wine Research*, 8(2), 75–80.

Lima, Tony. (2006). Price and quality in the California wine industry: an empirical investigation. *Journal of Wine Economics*, 1(2), 176–190.

Shrout, P.E. and J.L. Fleiss. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2), 420–428.

Thach, L. (2008). How American consumers select wine. *Wine Business Monthly* (June 2008), 66–71.